

DOCUMENT RESUME

ED 340 774

TM 018 030

AUTHOR Kober, Nancy
TITLE The Role and Impact of Chapter 1, ESEA, Evaluation and Assessment Practices. Part I: The Legal and Regulatory Framework for Chapter 1 Evaluation and Assessment; Part II: General Findings from Background Research; Part III: Case Studies; Part IV: Federal Policy Options.
SPONS AGENCY Congress of the U.S., Washington, D.C. Office of Technology Assessment.
PUB DATE Jun 91
NOTE 131p.; Contractor report prepared for the Office of Technology Assessment titled "Testing in American Schools: Asking the Right Questions." For related document, see TM 018 025.
PUB TYPE Reports - Evaluative/Feasibility (142)
EDRS PRICE MF01/PC06 Plus Postage.
DESCRIPTORS *Compensatory Education; Economically Disadvantaged; *Educational Assessment; Educational Finance; Educational History; Educationally Disadvantaged; Educational Policy; Elementary Secondary Education; Federal Legislation; *Federal Programs; *Program Evaluation; School Districts; *Supplementary Education
IDENTIFIERS *Education Consolidation Improvement Act Chapter 1; Elementary Secondary Education Act; Hawkins Stafford Act 1988

ABSTRACT

Chapter 1, Title I of the Elementary and Secondary Education Act of 1965 (ESEA) is the largest Federal program aiding elementary schools and secondary schools. Some 51,000 schools, including over 75% of the nation's elementary schools, receive Chapter 1 money, which they use to provide supplementary educational services to low-achieving students in poor neighborhoods. Evaluation requirements have been part of the authorizing law since it was first enacted. The 1988 amendments that govern the program today continue evaluation requirements. Chapter 1 offers an example of the way that Federal policy affects state and local decisions about evaluation and instruction. This study presents an overview of the role of evaluation in Chapter 1, a brief history of the evaluation requirements under Chapter 1, a description of the current evaluation and program improvement provisions, and an analysis of the cumulative effect of the Federal framework. This report includes the following parts: (1) the legal and regulatory framework for Chapter 1 evaluation and assessment; (2) general findings from background research; (3) case studies of the impact of Chapter 1 evaluation and assessment requirements on two local school districts; and (4) Federal policy options. Fifty-eight endnotes are listed. (SLD)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED340774

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as
received from the person or organization
originating it.

☐ Minor changes have been made to improve
reproduction quality.

• Points of view or opinions stated in this docu-
ment do not necessarily represent official
OERI position or policy.

The Role and Impact of Chapter 1, ESEA, Evaluation and Assessment Practices

Part I: The Legal and Regulatory Framework for Chapter 1 Evaluation and Assessment

Part II: General Findings from Background Research

Part III: Case Studies

Part IV: Federal Policy Options

**Prepared by Nancy Kober for the Office of Technology Assessment
June, 1991**

This contractor document was prepared for the OTA assessment entitled *Testing in American Schools: Asking the Right Questions*. It is being made available because it contains much useful information beyond that used in the OTA report. However, it is not endorsed by OTA, nor has it been reviewed by the Technology Assessment Board. References to it should cite the contractor, not OTA, as the author.

BEST COPY AVAILABLE

Tm 018030

PART I: THE LEGAL AND REGULATORY FRAMEWORK FOR CHAPTER 1 EVALUATION AND ASSESSMENT

INTRODUCTION

Chapter 1, Title I of the Elementary and Secondary Education Act of 1965 (ESEA) is the largest Federal program aiding elementary and secondary schools. With an appropriation of \$6.2 billion for fiscal year 1991,¹ Chapter 1 channels funds to almost every school district in the country. Some 51,000 schools, including over 75% of the nation's elementary schools, receive Chapter 1 dollars, which they use to provide supplementary educational services, primarily in reading and mathematics, to low-achieving children living in poor neighborhoods. Currently the program serves about 5 million children in preschool through twelfth grade.

Evaluation requirements have been part of the Federal authorizing law since 1965, when the program was first enacted as Title I of ESEA.² Although these evaluation requirements have been scrutinized and changed with every major rewrite of the law, their purposes have remained much the same: to determine the effectiveness of Title I/Chapter 1 in improving the education of disadvantaged children; to instill local accountability for Federal funds; and to provide information that state and local decisionmakers can use to assess and alter programs. This last purpose, which had received relatively less emphasis during most of the program's history and which was abandoned altogether as a statutory requirement from 1981 through 1983, took on new prominence in 1988, with passage of the most recent reauthorization of the Chapter 1 law. In a set of amendments known as "program improvement," the 1988 law, which governs the program today, requires Chapter 1 schools that cannot demonstrate

achievement gains among participating children to modify their local programs. The interaction of these new Federal program improvement provisions with the continuing evaluation requirements have made Chapter 1 assessment more consequential, controversial, and widely discussed than ever before.

Because of its broad reach and twenty-six year history of Federal guidance, Chapter 1 offers an unparalleled case study of how Federal policy concerning educational assessment affects state and local decisions about evaluation and instructional practices. As a starting point for such a study, this part: (1) presents an overview of the role of evaluation in Chapter 1; (2) provides a brief history of the Title I/Chapter 1 evaluation requirements contained in the Federal legal and regulatory framework; (3) describes in detail the current evaluation and program improvement provisions; and (4) analyzes the cumulative effect of the Federal framework. For purposes of this analysis, the Federal framework is defined to include the current Chapter 1 statute,³ regulations,⁴ and non-regulatory written guidance, particularly the guidance in the U.S. Department of Education's Chapter 1 Policy Manual.⁵

OVERVIEW OF THE ROLE OF EVALUATION AND ASSESSMENT IN CHAPTER 1 Federal, State and Local Responsibilities

Chapter 1 and its predecessor Title I divide program responsibilities among the Federal government, state educational agencies (SEAs), and local school districts ("local educational agencies" or LEAs in Federal parlance). School districts are responsible for designing and carrying out Chapter 1 programs that meet the special needs of disadvantaged children. The law affords local districts considerable flexibility to make

decisions about grade levels, subject areas, kinds of services, teaching methods, classroom settings, and types of staff funded by Chapter 1. States are charged with distributing funds to school districts, approving local applications, monitoring local projects for compliance, providing technical assistance and other statewide services, and running three special state programs. The Federal government allocates funds to states and also establishes, interprets, and enforces requirements governing the use of Federal dollars. Most of these requirements pertain to program inputs--ensuring that Chapter 1 reaches the neediest schools and students, does not supplant state and local resources, and operates in a fiscally responsible way. The evaluation requirements are among the few requirements that directly address program outcomes.

Evaluation responsibilities under Chapter 1 are similarly divided among the levels of governments. The evaluation of the effects of Chapter 1 on student achievement begins at the project level--usually the school building level. Evaluation results from individual projects are collected, analyzed, and summarized by the school district. Each LEA reports its findings to the SEA, which uses the local data to develop a statewide report that the state, in turn, transmits to the U.S. Department of Education (ED). The Congress, through the statute, and ED, through regulations and other written guidance, set standards for states and LEAs to follow in evaluating and measuring progress of Chapter 1 students. The Department is also responsible for compiling the state reports and transmitting to Congress a national summary of Chapter 1 effectiveness.

Importance of Standardized Tests

As Chapter 1 evaluation has evolved over the years, it has become nearly

synonymous with norm-referenced standardized testing. Since the first attempts to create a uniform evaluation and reporting system in the mid-1970s, the Department has looked toward norm-referenced standardized test scores as an available, straight-forward, and economical way of providing national data about how well children served by Chapter 1 are achieving and satisfying Congressional demands for a national picture of Chapter 1 effectiveness.

The current system, the Title I/Chapter 1 Evaluation and Reporting System (TIERS), became operative in 1979. TIERS aggregates state and local data showing changes in Chapter 1 students' achievement in reading, mathematics, and language arts and also compiles demographic data about Chapter 1 participants. So that this data can be aggregated for the whole nation, states and local districts must report gains and losses in achievement in terms of Normal Curve Equivalents (NCEs), a statistically-based score developed specifically for Title I that resembles a percentile score but has properties that enable it to be used to compute group statistics, combine data from different norm-referenced tests, and evaluate gains over time.⁶ (An NCE gain reflects an improvement in position relative to other students.⁷) Although Chapter 1 law does not explicitly require norm-referenced testing, the use of NCEs for well over a decade has given a Federal imprimatur to norm-referenced tests (NRTs) or other tests whose scores can be equated with national norms and aggregated.

This desire for aggregated national data has been a driving force behind Congressional and Departmental decisions about evaluation. At the same time, however, Federal legislators, evaluators and program administrators have acknowledged

the limitations of test scores in providing a full and accurate picture of Chapter 1's accomplishments.

The Many Uses of Standardized Tests in Chapter 1 Assessment

Producing data for national evaluation reports is only one of several uses of standardized tests in Chapter 1. Under the current Chapter 1 framework, LEAs are required, encouraged, or permitted to use tests for all the following decisions:

- identifying which children are eligible for Chapter 1 services and establishing a "cutoff score" to decide which children will receive services;
- assessing the educational needs of Chapter 1 children;
- determining the base level of achievement of Chapter 1 children before receiving services (the "pre-test") and assessing the level of achievement of Chapter 1 children after receiving services (the "post-test"), in order to calculate the change data required for national evaluations;
- deciding whether schools with high proportions of low-achieving children should be selected as sites for Chapter 1 programs, ahead of schools with high poverty;
- allocating funds to individual schools;
- establishing goals for schoolwide projects (high poverty schools that are allowed greater flexibility in the use of Chapter 1 funds, in exchange for increased accountability) and assessing student needs in schoolwide projects;
- determining whether schoolwide projects can be continued beyond their initial three-year project period;
- annually reviewing the effectiveness of Chapter 1 programs at the school building level for purposes of program improvement;
- deciding which schools to target for program improvement;
- determining when a school no longer needs program improvement;

- identifying which individual students have been in the program for more than two years without making sufficient progress; and
- assessing the individual program needs of students that have participated for more than two years.

In addition, the Congress and the Department of Education use standardized test data accumulated from state and local evaluations for a variety of purposes:

- justifying continued appropriations and authorizations;
- weighing major policy changes in the program;
- targeting states and districts for Federal monitoring and audits; and
- along with other assessment data, carrying out Congressionally-mandated studies of the program.

Although it is not uncommon for Chapter 1 teachers, classroom teachers, and administrators to employ other types of assessment to make decisions about individual Chapter 1 students and projects, these alternative forms are not entrenched in the program's basic operations in the same way that norm-referenced, standardized tests are, and are not usually considered by local people to be part of the "formal" Chapter 1 evaluation process. The description of the current legal framework below contains more detailed information about the uses of various forms of assessment in Chapter 1.

HISTORY OF LEGAL AND REGULATORY REQUIREMENTS

Balancing Tensions

The current Chapter 1 evaluation and program improvement requirements are an outgrowth of twenty-six years of political and programmatic decisions. This is not to say that the evaluation framework was built slowly and methodically, with components added as needs arose. Rather, Congress and the Department have changed the law and

regulations affecting evaluation many times and in very different ways, sometimes attaching far-reaching requirements, and once deleting whole sections of the law.

The number of times Congress has revisited the Chapter 1 evaluation requirements suggests how difficult it is to balance the many tensions inherent in the statute and the program. A tension has always existed in Title I/Chapter 1 between the need for Federal accountability on one hand, and the need for state and local flexibility on the other. And though the Federal government has a healthy appetite for "hard data", Members of Congress are acutely aware of the undesirable consequences of too much Federal burden and paperwork--lost instructional time and declining political support for Federal programs, to name a few. Choices about assessment instruments are also fraught with tension: on one hand, Congress needs clear and relatively uncomplicated national data to justify annual appropriations and periodic reauthorizations of Chapter 1; on the other hand, school districts need rich, meaningful, child-centered information on which to base day-to-day decisions about instructional methods and student selection.

Tension has also persisted between the program's anti-poverty focus and its achievement goals. While standardized achievement tests may serve as indicators of participants' educational progress in the basic subjects, they reveal little about progress toward other overt or implied program goals, such as redistributing resources to poor areas, mitigating the social effects of childhood poverty, building children's self esteem, and keeping students in school.

In a sense, then, the history of Chapter 1 can be viewed as a series of attempts by Congress and the executive branch to balance these inherent tensions, with varying degrees of success. Through the 1960s and 1970s, these attempts took the form of increased prescription and detail. In 1981, a dramatic loosening of Federal requirements occurred--to be followed, during the 1980s, by a gradual retightening.

1965 through 1974

The basic principle that Title I should be evaluated periodically was implanted in the law from the outset. The original 1965 Act contained a provision, sponsored by Senator Robert F. Kennedy, requiring that "effective procedures, including provision for appropriate objective measurements of educational achievement, will be adopted for evaluating at least annually the effectiveness of the program in meeting the special educational needs of educationally deprived children"⁸

The data that resulted from these locally-conducted and state-synthesized evaluations was, in the words of two researchers, "impossible to summarize . . . [and] far too divergent in terms of quality, quantity, and focus."⁹ These disappointing results were a perennial cause of frustration among policymakers and scholars, some of whom blamed the law for delegating responsibility for this important task to state departments of education and school districts,¹⁰ which often as not lacked the technical expertise to design quality assessments and the incentive to publicize less than glowing outcomes.

The results of national evaluations of Chapter 1 during the last 1960s and early 1970s also concerned many Federal policymakers, in that most of these evaluations did

not show strong achievement gains among Chapter 1 students. This concern also helped motivate policymakers to revise evaluation requirements.

The persistence of inadequate evaluations led to the adoption of a major evaluation section in the 1974 amendments to ESEA. These amendments marked the real beginning of the development of a formal, national system of Title I evaluation. The 1974 law required LEAs to evaluate their programs at least once every three years and to include objective measures of educational achievement in basic skills over at least a twelve-month period. SEAs were required to make periodic evaluation reports to the Office of Education. Most significantly, the law required the U.S. Office of Education¹¹ to develop and implement "standards for evaluation of program or project effectiveness in achieving the objectives of Title I."¹² As Congress envisioned it, this evaluation system would have several components: (1) "models for evaluations of all programs" conducted under Title I; (2) Federally-provided technical assistance to SEAs "to enable them to assist local education agencies in the development and application of a systematic evaluation of programs in accordance with the models"; (3) an annual evaluation report to Congress; and (4) a dissemination program to inform SEAs and LEAs about exemplary Title I projects.¹³

Spurred by this mandate from Congress, the Office of Education began designing models for interpreting test scores and calculating gains of Title I students (including NCEs), and developing standardized procedures for state and local reporting of achievement and student demographic data. Together these models and reporting standards formed the basis of TIERS, a system that, when fully implemented in 1979,

would make it possible for data from several thousand LEAs to be aggregated into a national summary of Title I's effectiveness.

To carry out the 1974 mandate for Federal technical assistance, the Commissioner of Education in 1976 established a national network of ten Technical Assistance Centers (TACs), which provided SEAs with specific services to improve the quality and use of evaluations and enhance the effectiveness of local programs. These centers continue to be an important resource in the Federal evaluation and assessment effort; TACs educate states and local districts about technical testing issues and interpretation and help them design appropriate assessments and program modifications.

1978 Amendments

It took several years, numerous research contracts, and countless communications with Congress for the Office of Education to develop TIERS. During this time, the Office wrestled with several technical issues regarding the relevance and quality of norms, conversion errors, and validity of estimates. Several reporting requirements and models emerged, including the annual testing cycles, technical requirements guiding states that wished to sample, and the popular Model A. (Model A uses test publishers' norms to approximate how much Chapter 1 children would be likely to progress in the absence of Chapter 1 services, then compares the approximations with the actual achievement of Chapter 1 children, expressed in NCEs). These reporting requirements and models laid the groundwork for the system as it exists today.

Although 40% of LEAs were using the TIERS models by 1977,¹⁴ TIERS was not fully operational before Congress began its next review of the Title I law. Reacting to

state and local concerns about implementation of TIERS, the 1978 amendments to ESEA added a requirement that the Office of Education consult with state and local evaluators in the design of TIERS. In keeping with the overall thrust of the 1978 amendments toward increased detail and precision, Congress made other clarifying changes in the law's evaluation requirements. It should be noted, however, that Congress was beginning to express some nervousness about the ultimate impact of a single, large, national evaluation system; both the House and Senate reports on the 1978 bills raised questions about unnecessary reporting burden and duplication and about the feasibility of meeting Federal needs for aggregated data and local needs for reliable information with a single system.¹⁵

Questions of burden and utility aside, TIERS soon achieved its goal of universal participation. School year 1979-80 was the first year that states were required to report data using the TIERS format and models, and was the only year before 1989 in which all SEAs complied.¹⁶

Education Consolidation and Improvement Act of 1981

Enacted in the anti-regulatory fervor of President Reagan's first term as part of a massive budget reduction bill, the 1981 Education Consolidation and Improvement Act (ECIA) repealed several provisions of Title I that were deemed to be too prescriptive and burdensome; redesignated Title I as Chapter 1; and reduced the statute to a shadow of its former length. Although the requirement that school districts evaluate their programs at least once every three years was preserved, gone were the directives that LEAs and SEAs use TIERS or any other standard evaluation procedures; that LEAs

employ evaluation results to improve their programs; and that SEAs collect evaluation data. As part of a general reduction in Federal regulatory power, ECIA also took away the requirement for the Department of Education to provide technical assistance concerning evaluation, and in fact explicitly prohibited the Secretary of Education from issuing regulations relating to "the details of . . . evaluating programs and projects by . . . local educational agencies".¹⁷

Despite this relaxation of rules, it appears that most states and LEAs made only guarded changes in their Chapter 1 programs and evaluation procedures. In the process of eliminating Federal mandates and heavy-handedness, ECIA had also done away with statutory language that spelled out options for program design and implementation and clarified state authority and local discretion. In the absence of detailed guidance, states and LEAs were generally confused and uncertain about how much latitude they really had and whether too much loosening would hurt disadvantaged children; most chose to play it safe.

Recognizing the confusion, Congress responded in 1983 by enacting a set of so-called technical amendments, which restored some, but not all, key provisions of the former Title I law. As regards evaluation, the technical amendments reinstated the former requirement that SEAs collect data and evaluate programs, changing the frequency of state evaluations from every three years (the TIERS requirement) to every two years. Rather than requiring SEAs to submit their evaluation reports to ED, the law instructed SEAs to "make public the results." In addition, SEAs were required to collect data on numbers, grades levels and race of participating children, which had been

collected under TIERS, and on age and gender, which had not.¹⁸ The requirement for LEAs to consider evaluation results in improving their programs was also restored, in slightly less prescriptive language. The mandated use of TIERS models and formats, however, was not restored.

Following enactment of the technical amendments, ED encouraged (but did not require) states to submit aggregatable data annually, continued sponsoring TAC services, and, in non-regulatory written guidance, noted that the TIERS models, although not mandatory, were "appropriate evaluation designs" and that states could require LEAs to use specific measures or evaluation designs.¹⁹ By 1987, 39 states had formally required the use of TIERS models by dint of their own state rulemaking powers; another seven reported that all their school districts used the models without a formal requirement. In addition, 36 states required annual submission of evaluation data. Altogether, about 90% of LEAs were using one of the standard TIERS models in 1987, and the same percentage submitted evaluations as often as they did under Title I.²⁰ By 1987, all states were also transmitting at least some data on participation.²¹

From this brief history of prior law emerge several pertinent themes:

- the emphasis throughout Title I/Chapter 1's history on "objective measures" of student achievement;
- the Congressional desire for a national, standardized system of aggregatable data;
- the influence and endurance of TIERS models and technical standards; and
- the interest of most states in maintaining an evaluation and reporting framework, even without a Federal requirement to do so.

KEY PROVISIONS OF THE CURRENT FEDERAL FRAMEWORK

Overview of 1988 Hawkins-Stafford Amendments

With the passage of P.L. 100-297, the Augustus F. Hawkins-Robert T. Stafford Elementary and Secondary School Improvement Amendments of 1988,²² Chapter 1 evaluation and assessment entered a new era. In a nutshell, the 1988 amendments:

- retained the existing provisions for LEAs to conduct evaluations every three years and for SEAs to submit results every two years;
- restored the requirement from the 1978 amendments that state and local evaluations conform to national standards and gave ED new regulatory authority to develop these standards, so that Congress could obtain "a national picture of how well Chapter 1 is accomplishing its intended goals of improving achievement";²³
- emphasized the need for Chapter 1 children to acquire more advanced analytical, reasoning, and problem-solving skills, in addition to basic skills;
- directed LEAs to develop "desired outcomes" for their Chapter 1 programs; in other words, to set goals against which progress could be measured;
- made the continuation of schoolwide projects dependent on the achievement gains of Chapter 1-eligible students in project schools; and
- instituted an interrelated set of "program improvement" provisions that, for the first time, put teeth in the longstanding requirement that LEAs use the results of evaluations to improve instruction; these provisions require schools in which the achievement of Chapter 1 children is stagnating or declining to develop plans for modification in conjunction with the district, and eventually with the state.

For many Members of Congress, the quality of Chapter 1 programs had become a recurrent concern. On both sides of the aisle, both chambers of Congress, and both ends of Pennsylvania Avenue, a feeling existed, not always articulated, that after more than two decades, Chapter 1 should be doing more than helping children make modest gains; it should be helping to close the wide gap that remained between them and their more

advantaged peers. (It bears noting that Chapter 1 students are by definition a very low-achieving population. According to ED's most recent summary report of achievement data, the average entering Chapter 1 student performs at the 22nd percentile in reading and the 26th percentile in mathematics.)²⁴ The answer, many Members of Congress felt, was to create a climate of higher expectations, in which programs that used the same instructional methods year after year with mediocre results, would be judged inadequate. As the House Report concluded, "Most Chapter 1 programs are producing satisfactory results, but the existence of truly exemplary programs suggests that more programs could do even better. Moreover, there is no reason why any Chapter 1 program should be operating poorly or why a program should 'give up' on any child who is making no progress."²⁵

Thus, the intent of the 1988 amendments was to move the Chapter 1 legal framework beyond its traditional emphasis on targeting and fiscal compliance, toward a new emphasis on program quality and greater accountability for student performance. As states and LEAs move into their third year of implementing the 1988 amendments, many are discovering an unanticipated consequence: intended or not, the net effect of the evaluation and program improvement requirements is to raise the stakes placed on local evaluation data and give even more primacy to the role of standardized tests in Chapter 1.

Description of Key Provisions

Evaluation and assessment issues are addressed in fifteen separate sections of the law. To administer the law, the Department of Education promulgated final regulations

on May 19, 1989, of which seventeen sections deal in some way with evaluation, assessment, and program improvement. Another 50 pages of the Department's Chapter 1 policy manual provide guidance on these same issues.

The policy manual is a key component of the Federal framework for Chapter 1. Published by ED in compliance with section 1436 of Hawkins-Stafford, the manual gathers under a single cover all pertinent requirements, explanations, opinions, statements, instructions, court decisions, and interpretations concerning Chapter 1. Its purpose is to help clarify and elaborate on the law and regulations and to answer common questions about Chapter 1 policy and procedures. According to the Department, LEAs may rely on the guidance in the manual unless it is inconsistent with their SEA's guidance; however, LEAs are free to develop alternative approaches consistent with the statute and regulations (p. i). Despite these directions, local Chapter 1 personnel tend to follow written Departmental guidance, such as that contained in the manual, as though it were the letter of the law; fear of Federal audit exceptions and misunderstanding about how much flexibility really exists in the law are some reasons why many local programs choose the play-it-safe route.

Following is a detailed, issue-by-issue description of the major evaluation and assessment provisions of the law, regulations, and policy manual.

Section 1012: Application and Desired Outcomes

The 1988 amendments require LEAs to "describe the desired outcomes for eligible children, in terms of basic and more advanced skills that all children are expected to master, which will be used as the basis of evaluating the program or project

as required by section 1019 . . .²⁶ The regulations impose the additional requirement that LEAs describe in their application how they will measure substantial progress toward meeting these outcomes (200.20)--regulatory language that strengthens the link between this section and the program improvement sections of the legislation.

The concept of desired outcomes is a new one in Chapter 1, and the regulations and policy manual seek to illuminate its meaning with additional interpretations and explanations. According to the regulations, desired outcomes are the goals an LEA sets to help children succeed in the regular school program, attain grade level proficiency, and improve achievement in basic and more advanced skills (200.6). (The language describing the three goals for desired outcomes is taken almost verbatim from the Chapter 1 law's overall statement of purpose.) The regulations and the policy manual further require that at a minimum, desired outcomes must be expressed in terms of aggregate performance and may be expressed using such indicators as criterion-referenced tests (CRTs), improved attendance, fewer retentions in grade, lower dropout rates, and writing samples (200.6 of the regulations and pp. 34 and 137 of the policy manual). The manual encourages LEAs to develop additional desired outcomes that are "more directly related to their Chapter 1 programs," while at the same time requiring all outcomes to be stated in measurable terms (p. 34). These additional specifications in the regulations and policy manual set the stage for subsequent provisions in the Federal framework that base program improvement decisions on progress toward desired outcomes.

Section 1014: Eligible Children

The 1988 amendments and accompanying regulations continue the requirement for LEAs to assess annually the educational needs within the district, as a basis for identifying all children eligible for Chapter 1. LEAs must select children with the greatest needs for services, the law says, using "educationally related objective criteria established by the local educational agency, which include written or oral testing instruments that are uniformly applied . . . throughout the local educational agency."²⁷ Thus, the law requires performance on tests (although not necessarily standardized, norm-referenced tests) to be among the criteria used in student selection.

The references in section 1014 to objective criteria and testing instruments were added in 1988 but reflect longstanding Title I/Chapter 1 practice, whereby children are selected on the basis of test scores, used alone or in conjunction with other measures such as teacher judgment. Most commonly, LEAs set a cutoff achievement score somewhere below the 50th percentile, and serve all the children who score less than the cutoff, with certain exceptions based on other indicators.

The regulations add a key clarification to the student selection provisions: namely, that limited English proficient (LEP) children may, in certain circumstances, be selected by measures other than standardized tests, such as teacher evaluations and language dominance tests (200.32). This exception for LEP students was first proposed in Congressional report language.

The policy manual amplifies the student selection procedures in several ways. First, the manual gives examples of information that an LEA could collect during the

needs assessment phase, including standardized tests, districtwide formal assessment results, student records, surveys of professional staff, and state or locally-mandated tests (pp. 61-62). Second, it makes clear that the needs assessment can be conducted without using norm-referenced tests (p. 124). Third, it clarifies that norm-referenced tests are just one example of objective criteria; others include CRTs, teacher assessments of performance, reading or math placement assessments, and other educational assessment instruments; and at the preschool level, readiness tests, diagnostic developmental assessments, and teacher observations. As regards LEP children, the manual notes that CRTs may also be used for student selection. Fourth, it interprets the law's prohibition against supplanting to mean that Chapter 1 funds cannot be used to test students for identification purposes; if the district's regular testing program does not include certain grades, then "alternative educationally related information" must be used to identify children in the remaining grades (pp. 61-65).

Section 1015: Schoolwide Projects.

Under the 1988 schoolwide project option, high poverty schools--75% or more low-income children is the threshold--may use Chapter 1 funds to upgrade their entire instructional program for all children, rather than to support discrete projects that benefit only Chapter 1 eligible children. Although the option had existed prior to 1988, most schools did not take advantage of it because of the stringent local matching requirements tied to its use. The Hawkins-Stafford amendments removed the fiscal strings, but substituted strict accountability measures that hold schoolwide projects to a higher standard of academic performance for Chapter 1 eligible children than other Chapter 1 projects. Toward this end, the 1988 amendments and accompanying

regulations require schoolwide projects to demonstrate, after three years of operation, that the achievement of Chapter 1 eligible children in the school exceeds either: (a) the average achievement of Chapter 1 children across the LEA; or (b) the average achievement of Chapter 1 children in the same school during the three years before the schoolwide project started. For secondary-level schoolwide projects, lower dropout rates, increased retention rates, or higher graduation rates will suffice, as long as achievement does not decline in comparison with the three prior years. Schoolwide projects must collect achievement and other assessment data annually.

The policy manual makes explicit several requirements that are implicit in the law. First, Chapter 1 eligible children must still be identified using the LEA's achievement cutoff score, so that the LEA can determine the school's allocation and decide whether the accountability measures have been met. Second, the program improvement standards that apply to other Chapter 1 schools also apply to schoolwide projects. Third, Chapter 1 eligible students in schoolwide projects must exceed the performance of their Chapter 1 peers in all the subjects addressed by the district's Chapter 1 program--in both reading and math, for example--for the schoolwide project to continue.

Section 1019: Evaluation

Basic statutory provisions -- The section 1019 provisions are the core of the Chapter 1 evaluation requirements. Many of the requirements in the 1988 amendments were picked up from prior Title I or Chapter 1 law: requirements for LEAs to evaluate the effectiveness of Chapter 1 programs and submit results to the State at least once

every three years; to follow national standards developed by the Department; to determine whether improved performance is sustained for more than one program year (the so-called "sustained gains" requirement). New provisions clarify that local evaluations must use objective measures of student achievement in basic and more advanced skills aggregated for the entire LEA; and that evaluations of programs serving only preschool, kindergarten, and first grade children are exempt from these evaluation requirements.

As in prior law, Hawkins-Stafford requires SEAs to conduct evaluations using local data and submit their findings to the Secretary of Education at least once every two years. In addition, SEAs must annually collect and submit demographic data on the race, age, gender, grade level and handicapping condition (the last being a new requirement) of participating children.

As under Title I/Chapter 1, the Secretary must submit an evaluation summary report to Congress every two years. A key provision of the 1988 amendments that closely parallels the 1978 TIERS mandate, charges the Department with developing national standards in consultation with state and local educational agencies. In fact, the law and the Congressional conference report language explicitly sanctioned the continued use of the existing TIERS standards.²⁸ Related provisions in section 1462 of Hawkins-Stafford and in a subsequent technical amendment²⁹ require the Department to sponsor two major national evaluations of Chapter 1: a national longitudinal study of the prolonged effects of Chapter 1 participation on children; and a national assessment of how the 1988 amendments are working.

By directing the Department to develop national standards, the Congress gave ED a license to regulate that was both far-ranging and uncharacteristically broad for the post-Reagan era of Federal education law. The Department responded by adding numerous criteria and clarifications--contained in a new subpart of the regulations (sections 200.80 through 200.89) and 29 pages of the policy manual--to the requirements in the statute. Whether the requirements and standards went beyond what Congress had envisioned, in range and prescriptiveness, is a question only the Congressional authors can answer.

Use of NRTs -- The basic standard for local evaluations established in the regulations assumes the use of norm-referenced tests. According to the regulations, LEAs must assess the achievement of Chapter 1 children in grades two through twelve in reading, mathematics, and language arts and compare it with an estimate of what the achievement of these children would have been in the absence of Chapter 1. Since LEAs have no true way of knowing what children's achievement would have been without Chapter 1, the regulations instruct LEAs to use the performance of children in a normed sample as a proxy. The policy manual reaffirms the intent that school districts use NRTs to assess achievement "so that Chapter 1 data from across the nation can be aggregated to determine the overall impact of the program." The manual further requires results to be reported in terms of NCEs (p. 122).

Although alternative tests are ostensibly allowed, the regulations make it difficult for LEAs to exercise that option. LEAs that want to use an alternative test, such as a state or locally-developed achievement test, must obtain the permission of the SEA and

the Secretary--a request that the regulations ensure will not be granted unless the assessment instrument produces results that can be expressed in NCEs and equated to national norms (200.83). A handful of States and local districts have requested permission to use their own state or local standardized achievement tests; some have been denied, and others are still being considered by the Department.³⁰

Annual reporting -- While section 200.35 of the regulations repeats the statutory language requiring LEAs to evaluate programs using national standards just once every three years, the program improvement section of the regulations (200.38) makes clear that in actual practice the formal evaluation must be conducted annually. This occurs because the Department has interpreted ambiguous statutory language to mean that the annual review that LEAs conduct for program improvement purposes must be based on the results of the section 1019 evaluation and must assess whether students in each of the LEA's Chapter 1 schools made NCE gains. The effect of these seemingly inconsistent provisions is discussed in more detail below, in conjunction with program improvement.

Similarly, although the regulations seem to permit LEAs to report the findings of their annual evaluations just once every three years, the Department decisively discourages this practice in the regulations and the policy manual. Under pre-ECIA Title I law, states were permitted to report evaluation data using sampling procedures, in which only one-third or one-half of the LEAs in the state reported results each year. A limited number of states took advantage of this provision, and for a variety of technical reasons, the Department viewed the validity of such samples with suspicion.³¹ By

establishing high technical standards for sampling, the regulations make it difficult for SEAs to take advantage of this option, the effect of which is to require all LEAs to report every year. It should be noted that prior to 1988, all states had gone with an annual reporting cycle, so the regulations should have no practical effect. As required by law, LEAs must report demographic data each year.

12-month testing cycle -- The law is silent regarding the testing cycle for local Chapter 1 evaluations. The regulations and policy manual fill this vacuum by requiring achievement to be assessed across a 12-month testing cycle--either spring-to-spring or fall-to-fall--and by explaining in detail how and when tests should be administered. The Department notes that a 12-month cycle seems to provide a more reliable picture of student achievement. For many LEAs that had been using a fall-to-spring testing cycle, this requirement has spelled significant change. In school year 1987-88, before the new requirements took effect, over 668,000 Chapter 1 reading students were tested on a fall-spring cycle, compared with the 1,045,104 children tested on an annual cycle. In that same year, depending upon grade level and subject, anywhere from 24 to 44 States were reporting at least some achievement information using a fall-spring cycle,³² and these States have had to revise their schedules.

Basic and more advanced skills -- The regulations and the policy manual elaborate at length upon appropriate procedures for measuring progress in more advanced, critical thinking skills, a provision added to the law to address concerns that Chapter 1 instruction too often consisted of rote drill in basic skills. Section 200.80 of the regulations specifies that more advanced skills ought to be measured in reading using

the "comprehension" subtest of a reading NRT, and in math using the "problems and applications" subtest of a math NRT. As with basic skills, the results should be reported annually. The policy manual clarifies the types of subtests that can be used and the circumstances under which NRT scores for advanced skills must be reported separately. The policy manual goes on to acknowledge that "many advanced skills are not measured at all, or well, by NRTs" and encourages LEAs to use additional measures of advanced skills in evaluating progress toward desired outcomes (p. 127). As discussed below in the section on programs improvement, a lack of progress on measures of advanced skills measures could have meaningful consequences for a Chapter 1 program.

Desired outcomes -- The regulations strengthen the link between the section 1019 evaluation and the annual program improvement review by requiring LEAs to evaluate the effectiveness of Chapter 1 using two bases: aggregate achievement (NCE scores) and the desired outcomes described in their applications (200.35). This requirement, which does not appear in the law, seems to be part of the Department's overall attempt to resolve statutory ambiguities about the relationship between the mandated local evaluation and the required annual program improvement review. In the regulations, the Department has concluded that the evaluation data (NCE scores) are a key component of the information used in the annual program improvement review. As is discussed in depth in the program improvement section below, a case can be made that this may not have been Congressional intent.

The policy manual further notes that each desired outcome must be evaluated, using appropriate methods, which do not have to be standardized tests. (Examples of

measures are listed above in the discussion of the section 1012 application provisions). Moreover, LEA must collect data on desired outcomes every year, although this information need be reported only every third year.

Early childhood programs -- This is another area in which the policy manual attempts to interpret possibly confusing statutory language. The law exempts Chapter 1 early childhood programs (preschool, kindergarten and first grade) from the requirement for local evaluations in section 1019. The law does not, however, mention any exemption for early childhood programs from the program improvement requirements.

The policy manual states that early childhood programs are exempt only from the national standards for aggregatable data and that these programs should be evaluated in terms of desired outcomes. The manual cautions that NRTs "may not be appropriate for validly and reliably assessing the performance of young children; such tests should never be used as the sole measure . . ." The manual suggests developmental checklists, readiness tests, observational scales, teacher ratings, skills mastery checklists, attendance, and retention records as other possible measures (p. 144).

Progress in the regular program -- The regulations require LEAs to review the progress of Chapter 1 children in the regular instructional program (in other words, in the mainstreamed classroom with a regular district teacher). Although success in the regular program is one of the three basic purposes of Chapter 1 outlined in the first section of the statute, the requirement to evaluate progress in this area is not part of the law. ED defends imposing this requirement through regulation by noting that "gains reflected by achievement testing may not translate into improved performance in the

regular classrooms. Conversely progress in the regular program may not be reflected by achievement testing (p. 131)." According to the manual, success in the regular program ideally would be measured using multiple measures and using the same benchmarks of academic progress the LEA applies to other children. These measures may include additional testing in other subjects related to Chapter 1, although achievement test data is not required to assess progress in the regular program (p. 132).

Sustained gains -- The requirement that LEAs assess, at least every third year, whether performance is sustained over more than one program year has been part of the legal framework for some time. The provision addresses a problem inherent in a fall-spring testing cycle--the tendency of children to lose ground during the summer. Researchers have long recognized the likelihood of fall-spring testing producing higher gains than annual testing. The policy manual reiterates language contained in the Congressional conference report, which clarifies that LEAs may assess sustained gains by sampling or a carefully designed study, rather than tracking all children every year.³³ The manual states that NRTs are not the only means for measuring sustained effects; however, it states, the method used should match the method the LEA uses to measure improved performance in Chapter 1 (p. 129). For all intents and purposes, LEAs are therefore limited to NRTs. The manual also contains additional rules regarding the design, technical standards, and methodology for assessing sustained gains.

Technical standards for evaluation -- The policy manual establishes numerous other policies on a range of testing issues: which test editions and instruments are permissible; at what time of year tests must be administered; how to match pre-test and

post-test scores in schools with high student mobility; how to develop a sampling plan; and which professional technical standards apply to Chapter 1 evaluations.

Sections 1020 and 1021: Program Improvement

State program improvement plans -- The new program improvement requirements in the Hawkins-Stafford legislation vest significant responsibilities with SEAs. States are called upon to set minimum standards of performance, establish timetables for the various stages of the program improvement process, develop joint plans with schools that have reached the second stage of the process, and dispense Federal aid from a special fund and other types of assistance to schools that need help.

As a starting point for implementing the new provisions, the 1988 law directed each state to draw up a program improvement plan, in consultation with a statewide committee of teachers, administrators, parents and other practitioners, that served to notify LEAs and other interested parties of the standards and procedures that would govern critical program improvement determinations.

These plans (which are not the same as the joint plans the state develops with individual schools) were required by law to contain the objective measures and standards the state would use to assess aggregate performance under the program improvement review that each LEA must do annually. The regulations made a significant addition to this provision by requiring the plan to also contain the objectives and measures the state would use to assess "substantial progress toward meeting desired outcomes" (200.37). In addition, the regulations said, the SEA could establish standards in the plan for helping Chapter 1 children succeed in the regular program, attain grade-level proficiency, and

improve achievement in basic and more advanced skills. The law did not contain this provision, but did give states another option that was not mentioned in the regulations: the discretion to include in the plan state policies governing the section 1019 evaluation.

The rest of the regulatory language accompanying section 1020 closely paralleled the law. Plans must include: a timetable for key phases of the local program improvement process; the means the state would use to develop joint plans with LEAs; and the types of assistance that would be provided to schools that were designated for improvement.

The change in the regulations to include standards for desired outcomes reinforces the connection between the 1019 evaluation and the LEA annual program improvement review and is consistent with the Department's overall interpretation of the interdependence between these two processes.

The policy manual contained some key clarifications and recommendations about standards and measures. States could either set standards applicable to all LEAs in the state, or allow LEAs to develop their own, said the manual. In either case, the manual encouraged SEAs and LEAs to set standards higher than the minimum required by law--i.e., no gain or a decline in aggregate performance--to determine which schools needed program improvement (p. 163). All states have completed their plans, and the majority have elected to set the minimum standard allowed by regulation--that any gain over zero NCEs is sufficient to keep LEAs out of the program improvement process. Some states have also set standards by which LEAs should measure desired outcomes.³⁴

Local review and school identification -- Section 1021 of Hawkins-Stafford sets forth the program improvement requirements for LEAs, schools, and students. Because the Department has interpreted some ambiguous language in this section to bolster its case for NRTs and the other national evaluation standards, the exact statutory language bears repeating. Subsection (a)(1) requires LEAs to conduct an annual review of Chapter 1's effectiveness "in improving student performance for which purpose the local educational agency shall use **outcomes developed pursuant to section 1012 and subsection (b) of this section**" (emphasis added).³⁵ The law goes on to say that LEAs must determine whether improved performance is sustained over a period of more than one program year (the regulations change this to 12 months); must use the results of the annual review and of the section 1019 evaluation in the program improvement efforts required later in the section; and must annually assess parental involvement.

The highlighted reference to "outcomes developed pursuant to section 1012" is straightforward enough: section 1012 is the local application section, and the outcomes are the LEA's desired outcomes. What Congress meant by the reference to subsection (b), however, is unclear. Subsection (b) of section 1021 is a very long subsection that defines the circumstances under which schools are subject to program improvement; lays out the processes a school must undergo during the first stage of program improvement; and describes the contents of a local program improvement plan. The word "outcomes" is used twice in subsection (b): the first time in reference to the desired outcomes in the local application; the second to the LEA's "stated program outcomes"--presumably the same as the desired outcomes. The vagueness of the language leaves a great deal of

room for interpretation and clarification in the regulations, and the Department has filled in the spaces.

Before discussing ED's interpretation, it is beneficial to review the criteria laid out in subsection (b) for determining which schools are subject to the program improvement process. The law states that each school which does not "show substantial progress toward meeting the desired outcomes . . . or shows no improvement or a decline in aggregate performance of children served under this chapter for one school year as assessed by measures developed pursuant to section 1019(a) or subsection (a)" (emphasis added) must develop and implement a plan to improve its program.³⁶

Here again, in the highlighted language, the statute contains an ambiguous cross reference to another subsection. The mention of section 1019 refers plainly enough to the objective measures the LEA uses in its mandated evaluation. But the reference to subsection (a) is vague. The multiple cross references create a baffling, circular logic that might be paraphrased like so: subsection (a) requires LEAs to review their programs using the desired outcomes in their applications or outcomes developed pursuant to subsection (b); and subsection (b) permits LEAs to assess aggregate performance using the mandated evaluation measures or measures developed pursuant to subsection (a). Despite the obscurity, one could still reasonably infer that Congress wanted LEAs to have two options for measuring aggregate performance: the objective measures used in the formal evaluation--in other words, norm-referenced test scores--or another, less formal measure developed specifically for purposes of the annual review. Just what this measure is intended to be is difficult to surmise. It does not seem to be the same as the

measures of desired outcomes, because those are addressed separately in subsection (b).

The regulations replace the ambiguities in the statute with rather precise language that strengthens the Department's case for annual formal evaluations and that installs NCE data at the center of the program improvement process. First, the regulations state that the annual local review must assess the program's effectiveness in terms of both desired outcomes and aggregate performance, and that a failure to show progress under either measure qualifies a school for program improvement procedures. This seems to be a reasonable interpretation of the legislative language. When coupled with the earlier guidance in the policy manual that allows an LEA to express its desired outcomes solely in terms of achievement (p. 33), the result is that an LEA could judge a program solely by norm-referenced test scores.

Second, the Department ties the measure of aggregate performance back to the national standards--i.e., in terms of NCE gains or losses. The regulations conclude that no improvement or a decline in aggregate performance has occurred when the Chapter 1 children in a school, in the aggregate, "fail to make gains beyond that which they would be expected to make in the absence of the additional help the program provided" (200.38). The policy manual explicitly defines the minimum standard for no improvement or a decline as "a positive NCE gain measured on an annual testing cycle" (p. 154). Together these Departmental assertions about aggregate performance return, full circle, to the national standards and constitute the final flagstone in the Department's case for an annual formal evaluation that complies with ED standards. The logic appears to proceed like this: LEAs must annually review program effectiveness

in terms of (a) progress toward desired outcomes; and (b) improvements in aggregate performance. The measures used in the formal evaluation are the only way--or at least one way, if the reference to subsection (a) is assumed to mean something--of assessing aggregate performance. Therefore, the 1019 evaluation must really be done annually, not every three years, in order for LEAs to have results they can use in their annual review.

There is certainly support in the statute for the Department's insistence on NRT measures as a means for assessing aggregate performance--the multiple cross-references back to section 1019(a), for example. The statute is far from crystal clear, however, and it is difficult to say with certainty what Congress intended. For example, language in subsection 1021(b) of the law might be construed as establishing two alternatives for measuring aggregate performance (although the law gives no specifics about what other alternative was intended). One could argue that Congress intended the annual program improvement review to be a quite different process from the formal evaluation required by section 1019, and one could even make a case that Congress wanted LEAs to be able to use a measure in addition to NRTs to determine which schools must undergo program improvement. Alternately, one could argue that even if the "measures developed pursuant to section 1019(a)" are the **only** legal way to assess aggregate performance, the statute would still permit an LEA to administer a NRT annually without following all the other requirements of the national standards, and to do a full-scale, formal 1019 evaluation every three years.

The regulations also clarify that an LEA must determine a school's aggregate performance only in the subject area that is the primary focus of the school's Chapter 1 program, unless two or more areas are given relatively equal emphasis (200.38). If an LEA has expressed its desired outcomes solely in terms of aggregate performance, then the policy manual requires the LEA to make program improvement judgments using two achievement measures: the primary subject area addressed in its Chapter 1 program would be constitute the basis for measuring aggregate performance, and the secondary subject area would be the basis for determining progress toward desired outcomes (p. 157). Thus, a school's project could quite possibly end up in program improvement solely on the basis of standardized test scores in reading and mathematics.

The policy manual contains elaborate guidance about how an LEA identifies schools for program improvement. Among those interpretations with the greatest relevance to assessment issues are the provisions in the manual that:

- Clarify that a decline in either basic or advanced skills is adequate grounds for triggering program improvement.
- Urge LEAs to set desired outcomes commensurate with the goals fixed for other children in the state; for example, if the state establishes mastery goals for a state test, then those same goals should apply to Chapter 1 children (p. 150).
- Give examples of appropriate measures of desired outcomes, including gains in NCEs; percentage of children that do not attain the district's specified skill levels; passing rates on state competency tests; percentage of children retained in grade; and dropout rates (p. 151).
- Clarify that for program improvement determinations, achievement data should be aggregated by subject area for all grades in a single school (p. 153).

- Prohibit LEAs from applying "error bands" to standardized test scores (+ or -3 NCEs, for instance) when making decisions about which schools must undergo program improvement (p. 154). (If error bands were applied, far fewer schools would have to undergo the program improvement process, even those whose true NCE scores were lower than indicated without the use of error bands.
- Delineate situations in which LEAs with small programs may use median NCE scores instead of the mean for program improvement determinations (p. 154).
- Explain that desired outcomes must be the same for all the schools and students in the LEA (p. 155).
- Answer questions about how to aggregate gains for students who transfer schools; whether to include children who enter late in the year; and when to include LEP students' scores.

Stages of program improvement -- The law sets up two stages of program improvement: the local plan stage and the joint plan stage. In the first stage, LEAs must develop and implement a program improvement plan for every school in the district that does not show substantial progress toward the desired outcomes or shows a decline or no improvement in aggregate performance. The plan, developed in coordination with the school, must describe how the LEA will make changes in the school's Chapter 1 program that are likely to improve the performance of participating children. The law describes at length some of the strategies that might be used to modify programs. LEAs may apply to the state for special program improvement funds authorized by section 1405 of the Federal law. Schools with fewer than 10 students are exempted from the program improvement requirements.

If the plan is in effect for the time set forth in the state guidelines and the aggregate performance of Chapter 1 children still does not improve, then the second

stage commences. At this stage, the LEA and the SEA together develop a joint program improvement plan, in consultation with school staff. The SEA must provide assistance, financial and otherwise, to schools with joint plans.

If, after one year under the joint plan, the school still does not meet the program improvement performance standards, then the SEA and LEA annually review and revise the plan until improved performance is sustained for more than one year.

The regulations track the law fairly closely on these issues, although they do make some clarifications pertinent to assessment. If data becomes available prior to implementation of a local plan that demonstrates that the school has met the LEA's standards after all, then the school does not have to go through with the plan. The regulations also set maximum timelines for implementation of local and joint plans, but encourage schools to move forward expeditiously with at least a part of the plan (200.38).

The policy manual clarifies that if the SEA and LEA cannot agree on a joint plan, the SEA could ultimately refuse to approve the part of an LEA's application related to the school(s) in question; however, LEAs and SEAs should strive to avoid this situation (p. 164).

Local conditions -- The law and regulations list five "local conditions" that could affect the performance of Chapter 1 students and that LEAs and SEAs must take into consideration when carrying out the program improvement provisions. The conditions are: (1) student mobility; (2) the extent of educational deprivation among a school's Chapter 1 children; (3) difficulties involved in dealing with older children in Chapter 1 secondary school programs; (4) whether indicators other than improved achievement

show positive effects for Chapter 1 children; and (5) whether changes in the review cycle, the measurement instrument or other "measurement-related phenomena" have rendered results invalid or unreliable. The regulations state that these local conditions may be considered at any point in the program improvement process.

The legal and regulatory framework do not definitively resolve the question of whether local conditions may be used to exempt a school from program improvement. The policy manual suggests that if the condition is something the LEA knew about beforehand, the district should address the situation when it allocates Chapter 1 funds among schools and may not later use the condition to excuse a school from program improvement. When unforeseen local conditions occur during the school year, the "the LEA may take that local condition into account" in determining whether program improvement is needed (p. 159). This language suggests that exemptions may be made in certain circumstances, such as a large increase in student mobility.

The manual does address in some detail what happens when an LEA uses different tests for different grades. If, after careful examination, the LEA determines that the results are invalid, even if the tests appear to be adequately equated, the LEA may claim a local condition to invalidate the results, but only once; after that, the use of the equated test must be discontinued (p. 160).

Student program improvement – Using information from the mandated evaluation in section 1019 and the annual program review, LEAs must identify individual students who have been in Chapter 1 for a year without making substantial progress and must consider modifying the program to serve them better. For students who remain in

the program for more than two years, the LEA must conduct a thorough assessment of their educational needs. These provisions are quite similar in law and regulation.

Attempting to resolve some ambiguities in the statute, the policy manual makes clear that student program improvement, like school program improvement, is determined by examining NCE gains and that student program improvement applies even when the school has not been singled out for program improvement (p. 162). The Department points to the existence of the individual student program improvement procedures as another reason why Congress must have intended annual evaluations using standardized tests.³⁷

Program improvement funds -- Section 1405 of the law authorizes appropriations to enable states to provide direct educational services to schools implementing program improvement plans. In years when no money is appropriated for this fund, SEAs still have to implement, at a minimum, the requirements for joint plans and carry out as much of their remaining duties as is practical (section 1020). The regulations direct SEAs to follow the progress of schools in the first stage of program improvement, even if there is no funding. In fact, Congress has appropriated these funds annually since they have been authorized.

Other sections -- Several other sections of law contain provisions relevant to evaluation and assessment. In brief, they are as follows:

- Section 1011 allows funds to be used for evaluation of Chapter 1 projects. (Section 200.88 of the regulations clarifies that this does not include test development, establishment of norms, or implementation of a state or districtwide testing program.)

- Section 1013 permits an LEA, under certain very limited circumstances, to serve schools with substantially higher proportions of educationally deprived children before it serves schools with higher concentrations of low-income children.
- Section 1016 allows LEAs to use Chapter 1 funds to provide evaluation information to parents. [The policy manual notes that LEAs should provide parents with evaluation data on a timely and continuing basis (p. 77).] Section 1016 also requires districts to annually assess the effectiveness of their parental involvement programs. According to the report accompanying the Senate bill, the effectiveness of parental involvement may be assessed by measures of student outcomes, such as increased attendance, or by measures of parent input, such as time parents spend with children on school-related matters or evidence of skills parents gain from the project.³⁸
- Section 1202 sets forth the standards for evaluating the state-administered Chapter 1 migrant education program. Migrant programs should be evaluated in terms of their effectiveness in achieving their stated goals; "objective measurements of educational achievement in basic skills" must be included.
- Section 1222 requires recipients of funds under the special Chapter 1 program for handicapped children to "collect and maintain such evaluations and assessments as may be necessary to demonstrate that the programs were beneficial to the children served."
- Section 1242 establishes the policy for evaluating state-administered programs for neglected and delinquent children. These annual evaluations must determine the impact of the program on children's ability to maintain and improve educational achievement; maintain school credit; and make the transition to a regular school program.
- Section 1435 directs the Secretary of Education to develop national standards for local evaluation (which may by law be based on the TIERS system) and to notify SEAs and LEAs in advance about these standards. This section also requires the Secretary to submit a biennial report to Congress summarizing state and local evaluation data.
- Section 1437 authorizes up to \$4 million annually for Federal research, evaluation, and technical assistance activities.
- Section 1451 gives states authority to issue regulations and charges them to work jointly with LEAs in program improvement.

- Section 1452 outlines the state responsibility to keep records and information required by ED for program evaluation.
- Section 1462 authorizes a national longitudinal study of the long-term effects of Chapter 1 participation, not only on achievement, but also on dropout rates, employment, delinquency, and other measures.

THE CUMULATIVE EFFECT OF THE FEDERAL FRAMEWORK

Interplay between the Law, Regulations, and Policy Manual

Together the Chapter 1 law, regulations, and policy manual form a framework for evaluation and assessment that is overwhelmingly detailed, somewhat confusing, internally inconsistent, and thoroughly reliant on norm-referenced standardized tests. It is also probable, as the discussion above tries to show, that the degree of reliance on norm-referenced tests was not always intentional. As the preceding descriptions illustrate, the Federal government is not speaking with a single voice as regards Chapter 1 assessment.

Each component of the framework makes a particular contribution to the overall effect. Although the 1988 statute is far more prescriptive than the ECIA law it replaced, it is still the least detailed component of the framework. It should be noted that the law set in motion an engine of detail by giving ED blanket (and uncharacteristic) authority to promulgate national standards through regulation. Had Congress known in advance the amount of prescription that would emerge in the national standards, one wonders whether it would have reconsidered its invitation to regulate.

The major problem with the legislation lies in its ambiguity--or its lack of a clear and coherent sense what Congress intended. The relationship between the section 1019 evaluation requirements and the section 1021 program improvement provisions is muddy

and spawns several unanswered questions. How often must LEAs conduct a formal evaluation? Which measures should they use to make program improvement determinations? What is the precise role of desired outcomes? In addition, the law is somewhat illogical in its requirements regarding the frequency of evaluations: every three years for LEAs, every two years for states, every year for program improvement.

This is not to say that the law was ill-conceived. Some ambiguities are likely to ensue when very different House and Senate versions are merged, when ongoing requirements are blended with brand new ones, and when Members of Congress attempt to accommodate different interests by accepting provisions that are not entirely consistent. To some extent, it is the nature of the legislative process.

The regulations possess a different set of problems. On the positive side, the regulations are internally consistent, make some important clarifications to the law, produce a logical and workable system for collecting large amounts of information, and adhere to high technical standards. On the negative side, one could argue that the Department may have overregulated in the national standards, and on most open questions, has come down on the side of rigidity and administrative burden. In addition, the regulations and policy manual include some questionable interpretations of Congressional intent--the annual evaluation requirement being the most prominent--and some significant additions above and beyond what the law says--the 12-month testing cycle, the requirement to measure success in the regular program, and the requirement to annually evaluate early childhood programs, for example. Some of these requirements

have been added for technically important reasons; the 12-month testing cycle is a good case.

This is not to say that the Department has a hidden, self-serving reason for collecting large amounts of data annually; rather, the Department's rationale seems to be that if a national system is mandated, it should produce data that is as valid and reliable as possible, and annual local evaluations, in the opinion of ED's evaluation staff, are more likely than other methods to furnish quality, aggregatable information.³⁹ Furthermore, an annual requirement seems to be a logical way to resolve some of the ambiguities and inconsistencies in the statute.

The policy manual adds extraordinary detail to the Federal framework. For better or worse, the manual reads like the work of testing technicians. In the interest of answering common questions, clearing up confusion, and pulling together all relevant interpretations, it imposes numerous requirements with no specific legal authority except the Department's general authority to develop national standards and properly administer the program. In fairness, it must be acknowledged that many of the interpretations contained in the manual have existed in Chapter 1 written and oral guidance for many years, and that perhaps it is first-time presentation of all this written guidance under one cover that makes it seem so imposing. Detail is not pernicious in and of itself; extensive guidance is often welcome by state and local people who must administer a program as large and complex as Chapter 1. However, the cumulative effect can be inhibiting, and states and localities that wish to try alternative forms of

assessment not described in the law or regulations will be mightily discouraged by a perusal of the policy manual.

It is informative to remember that to local practitioners, a Federal requirement is a Federal requirement, whether it comes from the law, the regulations, or the policy manual. Although much of the material in the policy manual is clearly identified as optional, this may be an inconsequential distinction to administrators who are subject to Federal enforcement and monitoring and who have lived through the consequences of Federal audits. The existence of state rules, on top of the Federal ones, can exacerbate the dampening effect.

Support for Commercially-Developed, Norm-Referenced Achievement Tests

While it appears on the surface that SEAs and LEAs have some flexibility to establish their own objective measures and standards, the Federal framework on the whole circumscribes flexibility. The framework seems to take for granted that the basic tool for measuring achievement of Chapter 1 children is the norm-referenced test. This is not surprising, given the history of Chapter 1 evaluation and the general reliance of the U.S. educational system on norm-referenced tests. In addition, testing issues have not been widely perceived as problems, or a main focus of debate, in Congressional reauthorization of Chapter 1, including even the most recent one.

With their numerous instructions on the technical aspects of testing--equating, sampling, subtests, norms, standard errors--the regulations and the policy manual appear to accept without question the premise of annual testing. Some provisions of the policy manual even recommend testing certain Chapter 1 children twice a year. Terms such as

"testing," "scores," "objective measures," "measuring instruments," and "aggregate performance" pepper the vocabulary of the law, regulations, and manual.

In addition, the law, regulations, and written guidance erect certain barriers to the use of alternative assessment and make it difficult to use sampling and other techniques that reduce testing burden. In many cases, the regulations require advance approval of alternatives and make clear that this approval—which had not been given any State or district at the time of writing—will not be bestowed liberally.

Limits of Standardized Testing

There is a certain amount of schizophrenia about testing in each component of the Federal framework. Even as the law, regulations and policy manual endorse NRTs, they are not averse to acknowledging their limitations and disavowing their usefulness for certain types of assessments.

The law, for example, exempts the state migrant, neglected and delinquent, and handicapped programs from the national evaluation standards, and notes that NRTs are inappropriate for assessing performance of LEP students and children in preschool, kindergarten, and first grade. To cite two other examples, the Department acknowledges that advanced skills may not be measured well by NRTs and that judgments about individual student progress made solely on the basis of test results may be inaccurate. Moreover, two of the five local conditions that States and LEAs must consider when carrying out program improvement, address the possibility of measurement errors: one is whether indicators other than achievement demonstrate positive effects of Chapter 1,

and the other is whether a change in measurement instrument or "other measurement-related phenomena" have rendered results invalid [section 1021(e)].

There is ample evidence that Federal legislators, evaluators, and program administrators are well aware of and, to varying degrees, concerned about the dominance of norm-referenced testing in Chapter 1 evaluation. National assessments of Chapter 1 in the mid-1970s and the mid-1980s, which Congress mandated as part of its preparation for major reauthorizations of the statute, have raised questions and caveats about the validity and limitations of Chapter 1 evaluation data. The most recent national assessment, completed in 1987, summarized and reanalyzed a range of existing data on Chapter 1 effectiveness in its second report, while cautioning that achievement test results are a limited and imperfect measure of Chapter 1 outcomes because they:

- do not measure the impact of other Chapter 1 services, such as supportive services;
- "measure achievement in areas that were never intended to be taught and will fail to document some of the skills that students learn as a result of participation";
- for a variety of reasons, "will generally underestimate the amount of program-related growth students have actually experienced;" and
- "do not reflect the long-term consequences of Chapter 1 program participation".⁴⁰

The 1987 national assessment raised several other concerns about Chapter 1 evaluation in general and TIERS data in particular. Chapter 1 is difficult to assess from a national perspective, the report noted, because local programs vary so widely in terms of the amount and intensity of instructional services, student selection procedures, administrative strategies, and other key characteristics that might affect student

outcomes.⁴¹ Moreover, because Chapter 1 is supposed to serve the lowest-achieving students in a school, it is difficult to find an appropriate comparison group.⁴² As regards TIERS, the study questioned its use of the entire distribution of scores as a comparison group; its basic assumption that without Chapter 1, students would maintain their same relative rank rather than falling farther behind; the accuracy of the state data on which it is based; and a number of other technical procedures. Other national studies have raised similar concerns.

Endorsements of Alternative Assessment

In many places throughout the framework, the Federal government permits and even encourages alternatives to NRTs. The measurement of desired outcomes opens up the broadest opportunities for the use of alternative assessment, albeit in conjunction with norm-referenced tests, to make more accurate judgments about student performance and the effects of Chapter 1 participation. Among the types of measures mentioned in the law, regulations, and policy manual are such commonly-used indicators as report card grades, dropout rates, teacher judgments, end of chapter tests, textbook levels, attendance rates, retention rates, and CRTs. The framework also cites less traditional forms of assessment, such as writing assessments, developmental checklists, observational scales, language dominance tests, and skill mastery checklists. As Parts II and III of this paper indicate, states and districts are using many of these options.

Conclusions

Considered together, the provisions described above form a rather daunting, externally-imposed framework that exhibits some symptoms of multiple personality

disorder. While the overall thrust of the framework is toward a uniform national testing policy for Chapter 1 evaluation and program improvement, there is significant acknowledgement of the limits of standardized, norm-referenced tests and some nods in the direction of alternative assessment (most often relatively conventional kinds). It is only logical that local administrators and teachers, operating under the umbrella of the Federal framework, would hesitate before venturing too far in the direction of alternative assessment.

PART II. GENERAL FINDINGS FROM BACKGROUND RESEARCH

INTRODUCTION

Research Methods

The contractor used a two-step process to learn more about how Chapter 1 evaluation requirements affect state and local practice and whether there is a role for alternative assessment in the program. First, the contractor did general background research, reviewing relevant reports and talking on the telephone or meeting with key people. Second, the contractor visited two school districts. This paper describes the broad findings of the background research. The findings of the case studies are discussed in Part III.

The background research was conducted first to help the contractor frame the issues and find the pressure points before doing the field work. The background research also turned up some valuable information in its own right about Chapter 1's effect on state and local practices and alternative assessment.

The contractor held in-depth telephone conversations with the following: Federal and Congressional staff; technical assistance center staff; state and local Chapter 1 coordinators; evaluators and researchers; staff of national associations representing school administrators, teachers, parents, state superintendents, civil rights advocates, and child advocates; and other persons knowledgeable about Chapter 1. The contractor also met with Congressional staff and Department of Education program managers and evaluators. Although the contractor worked from a list of general questions, the contractor did not do formal telephone surveys, choosing instead to let the persons voice their own concerns and help determine the key issues.

In addition, the contractor reviewed Department of Education publications, formal studies of Chapter 1, and books and reports about Chapter 1 evaluation and about assessment in general.

STATE AND LOCAL ATTITUDES ABOUT THE FEDERAL FRAMEWORK

High Interest

The 1988 amendments have made evaluation a front-burner issue in Chapter 1.

At the national level, the following activities have occurred roughly within the past year:

- The Department of Education has established an advisory panel on testing in Chapter 1, charged with making recommendations to the Secretary about improvements in the tests used to measure Chapter 1 or alternatives to the current testing procedures. As of this writing, the members of the panel had not been named. The panel is supposed to report within a year of its first meeting.
- The Department has held a series of regional meetings on Chapter 1 evaluation.
- A coalition of education organizations headed by the Council of Chief State School Officers is meeting to prepare for Chapter 1 reauthorization; evaluation and program improvement are among its priorities.
- Several groups, including the Chapter 1 state coordinators association and the Center for Law and Education, a parent advocacy group, are conducting surveys on Chapter 1 that include questions about evaluation and program improvement.

States are also focusing attention on Chapter 1 assessment. The California Department of Education, for example, has formed a working group on alternative assessment in Chapter 1.

Anxiety

It is the new program improvement and schoolwide project requirements, not the ongoing evaluation requirements, that have injected anxiety into the evaluation process.

For the first time, the evaluation data matters at the local level. The encouragement that states are receiving from ED and some Congressional leaders to set "high standards" (i.e., +3 NCEs instead of anything over zero) has exacerbated a climate in which schools live and die for a few percentile points, even as they dismiss them as statistically meaningless and educationally irrelevant.

Concerns about Testing

Although states and local districts are complying with Chapter 1 evaluation requirements, many express strong reservations about the usefulness of norm-referenced tests for assessing Chapter 1 children.

These are some widespread concerns:

- NRTs do not correlate well with Chapter 1 program goals, local curricula, higher order skills, or the modes of instruction being used in new reading and math programs.
- They provide teachers with little to no meaningful information about what a child is really learning, only how he or she compares with some larger group.
- They give just a pinhole glimpse of the positive effects of Chapter 1; teachers see many positive outcomes of Chapter 1 that are not captured by NCEs.
- They result in children being labeled when schools are moving away from that.
- They are inappropriate for early elementary children, children below the 20th percentile, and high school students who are years behind grade level.
- They do not take into account critical background characteristics that can affect student performance.

Another sizeable segment of Federal, state and local people believe that testing serves very useful purposes, including the following:

- They provide a check on bad or undemanding teachers, without which the needs of disadvantaged children cannot be protected.
- They "hold local districts' feet to the fire," as one state coordinator said.
- They are the only practical objective instrument that can satisfy the need for national, aggregated data.

Among those with concerns about norm-referenced standardized testing, the majority did not want to do away with norm-referenced testing in Chapter 1. State and local officials acknowledged its role as a limited national benchmark and its use in satisfying community and parent demands. As one local Chapter 1 coordinator said, "I know we'll always have to have standardized tests."

Cautiousness

SEAs and LEAs are being cautious about departing from past evaluation practices or even taking advantage of all the flexibility the law allows them. This cautiousness is characteristic of Chapter 1. Some Chapter 1 state and local directors have been with the program from the early years, and have seen flexibility come and go in the Federal statute. In addition, some Chapter 1 directors view their role as a Federally-stationed watchdog who keeps local principals and superintendents from going around the rules and protects the rights of disadvantaged children. Finally, some state and local Chapter 1 coordinators are accustomed more to a monitoring and compliance role than an educational and instructional one.

Most importantly, many states and school districts have been seasoned by Federal audits, and they know that test scores provide a clear, written audit trail.

The inconsistent and at times contradictory Federal framework can also inhibit creativity, as state and local program directors choose to fall back on past procedure when faced with uncertainty.

HOW SEAS AND LEAS HAVE IMPLEMENTED THE 1988 AMENDMENTS

Major Influences

In Chapter 1, state and local decisions are influenced not only by the law, regulations, and policy manual, but also by state rules and policies, Federal and state monitoring and audits, and oral guidance from Federal program managers, state officials, and other local Chapter 1 people.

State and local staff do not necessarily distinguish between what is required by the law and regulations, what is recommended by the policy manual, what is a state interpretation, and what is simply longstanding Chapter 1 practice. In fact, state and local officials do not always know the source of the requirements or interpretations they are following.

Delays

Many states and local districts have been slow in implementing the 1988 amendments. For example, many local districts are just now changing their evaluation cycles, setting desired outcomes, and learning about requirements they had overlooked. Several states have taken time to finalize state interpretations, especially program improvement standards, which have already been revised in some states. Not all of this delay is attributable to local foot-dragging. The ED final regulations came out more than a year after the law was passed, and the policy manual a year after that. Since

many of the provisions chart new waters, local districts did not want to proceed too far without guidance.

Evaluation and Testing

There is a general perception of Chapter 1 evaluation as a "hoop" that states and local districts must jump through to get Federal money. Because TIERS has been in existence for so long, it is taken for granted. Most states were already requiring a 12-month cycle and the use of Federal TIERS standards before 1988. The biggest change in regular Chapter 1 evaluation has occurred in districts that had used a fall-spring cycle. The most popular annual cycle for LEAS is a spring-spring one, using one of the commercial NRTs.

According to ED, all SEAs are complying with the national standards for evaluation. Some states--Connecticut, Missouri, New Jersey, and California--along with the Portland Public Schools and a consortium of districts in California, have requested permission to use a state or local test instead of a national NRT. Connecticut has been turned down unless it makes certain revisions; the California SEA has withdrawn its request because funding for state testing was cut. The others are still under review, with no specific timetable for ED to make its final decisions.

Program Improvement

All states are using NCEs or percentiles as the basic measure for program improvement, as required by ED. About two-thirds of the SEAs have set the minimum standard--any gain above zero--and this is after some states raised their standards between 1989-90 and 1990-91. The highest standard is +3 NCEs. There is evidence that

the setting of minimum standards is related to the amount of funds available for program improvement assistance; a state may not want more schools to be identified than it has money to help. Others have suggested that states do not want their LEAs to look bad or do not want the administrative burden of overseeing too many program improvement schools.

The use of NCEs to determine program improvement has heightened the level of controversy and anxiety about standardized tests in Chapter 1. Most schools perceive program improvement as a punishment.

Depending on whom one asks and when, the number of schools in program improvement varies from 10% to 15%. The Department is currently saying about 12% of Chapter 1 schools in the nation are in program improvement. The Department is looking carefully at the data on how many schools tested out of program improvement; according to ED representatives, a questionably high number of schools report enough improvement in a single year to allow them to move out of program improvement.

According to several studies, there seems to be little correlation between a state's NCE standard and the number of schools it has in program improvement; states with the minimum are about as likely to have many schools in program improvement as states with higher standards. Only a very few schools have reached the joint program improvement stage. This is because most states are allowing the maximum planning time of a school year for each stage.

Some states are using the local conditions section of the law to exempt schools from program improvement. The rigidity with which states interpret the local conditions

varies greatly; some are making several exceptions, while others do not interpret it as an exemption.

A majority of states addressed the issue of desired outcomes in their state program improvement plans, but most are leaving the choice of outcomes up to LEAs. Some states, such as Pennsylvania, Indiana, and Kentucky, have moved ahead aggressively with requiring LEAs to adopt desired outcomes.

Local districts are using a range of indicators for desired outcomes. They include CRTs, checklists, attendance and dropout rates, report card grades, book levels, graduation rates, reading fluency records, teacher assessments, "miscue" analyses, use of the library, participation in extracurricular activities, changes in school climate, and number of books read outside class. In the opinion of several state coordinators, some of the standards LEAs have set are too weak, or are not properly cast as a statement of goals. The way the law and regulations are structured, there is a disincentive to set desired outcomes. Some schools are in program improvement not because of a failure to meet aggregate performance standards but because they did not meet desired outcomes.

States vary in the degree of technical assistance they provide LEAs in program improvement. Some are conducting workshops and taking active roles in other ways.

EFFECTS OF THE FEDERAL FRAMEWORK ON PRACTICES

National Evaluation System: Amount, Time and Costs

Chapter 1 has helped engender an enormous system of state and local testing in which almost every Chapter 1 child is tested every year, and in some cases twice a year, to meet national evaluation requirements. Sometimes this testing is combined with

testing that fulfills state and local needs; other times Chapter 1 students take tests that would not have been given otherwise, and are often lose regular instructional time as these tests are administered.

According to the latest TIERS data, over 1.6 million Chapter 1 students (on a weighted basis) were tested in reading in school year 1987-88, and over 974,000 in mathematics.⁴³

How much time and money is spent on Chapter 1 testing is difficult to assess. ED has estimated that the 53 state educational agencies (the 50 States, the Districts of Columbia, the Bureau of Indian Affairs, and Puerto Rico) spend approximately 24,425 person-hours annually to provide both the achievement and the demographic data required for Chapter 1 reporting. ED has not attempted to estimate the burden for local respondents.⁴⁴

At the Federal level, the initial development and field testing of TIERS cost \$892,785; the annual costs to the Department to administer and report the Chapter 1 TIERS and demographic data is approximately \$165,000.⁴⁵ It is difficult to come up with any data about the costs of Chapter 1-related testing to states and local districts, because it is impossible to say with any certainty what they would do about testing requirements on their own, in the absence of Chapter 1.

Effect on Total Amount of Testing

The Federal Chapter 1 evaluation system did not spring up after the 1988 amendments; it has been 25 years in the making. Because SEAs and LEAs expanded their own testing programs during the same period, and because they often use the same

test for Federal, State and local purposes, it has become difficult if not impossible to sort out which entity is responsible for what degree of the total testing burden.

Some states and LEAs said they would do less testing without Chapter 1: test at every other grade, or in alternate years, or do less frequent testings. In many places, Chapter 1 forces testing to start at an earlier grade than SEAs or LEAs would like. One state coordinator said that in the absence of Chapter 1, his state would use only its state criterion-referenced instrument and no norm-referenced tests.

The Department of Education has discouraged sampling because of the poor data it received under sampling plans in the past.

Instruction

It is difficult to pin down the extent to which Chapter 1 drives instructional practices and curriculum because administrators and teachers are loathe to admit of anything that smacks of teaching to the test.

There is some suggestion, however, that time spent instructing children on skills likely to be on standardized tests is draining time from other, perhaps more meaningful, Chapter 1 instruction. Teaching skills that are likely to be tested may lead to an undue emphasis on the type of lower level skills that Congress has criticized Chapter 1 programs for stressing in the past. In addition, some local coordinators said that their teachers try to build Chapter 1 students' test-taking skills by such means as using teacher-made tests with similar formats to standardized tests.

There is also some suggestion that program improvement can distort instruction by focusing attention on the areas of the school that are not its greatest weaknesses.

Program Improvement

According to several state coordinators, program improvement has been a good incentive and a useful tool to get mediocre programs do a better job with Chapter 1 children. Several state and local coordinators said that the biggest benefit of program improvement was that it encouraged Chapter 1 and classroom teachers to talk to each other and improved coordination between Chapter 1 and the regular program. Some schools are using the program improvement process to completely restructure their programs.

State and local people expressed several concerns about the fairness, accuracy, and validity of program improvement decisions being based on NCEs. Here are some:

- The use of test scores for program improvement at times seems to produce an almost random selection of schools. It does not always target the schools with the greatest problems.
- In small Chapter 1 programs, such as those found in rural Vermont, the aggregation does not mean anything; one very low score can skew the results. Although the law exempts programs with 10 or less children, programs with more than 10 students enrolled can produce the same type of misleading results when scores are averaged.

Program improvement has created new administrative burdens at both the state and local level. A survey of state coordinators rated program improvement as the most burdensome category of provisions. (This survey was done in the first year of implementation, when the provisions were new and unfamiliar.) At the local level, the documentation of desired outcomes is also a new administrative burden that few had planned for.

Schoolwide Projects

The validity of standardized tests are of particular concern to schools with schoolwide projects. These schools have noted:

- that three years is not enough time for some results to show up in the form of improved test scores, yet that is what is required to continue a schoolwide project;
- that measurement error and other factors beyond their control can result in lower scores for some grades, which could force the school to discontinue its schoolwide project; and
- that the options for comparing their scores with the average of other Chapter 1 schools in the district or with their own past performance, to determine whether sufficient progress has been made, are not always fair ways to judge progress.

Burden

The Department of Education maintains, with some sound reasoning, that the requirement for an annual cycle in the 1989 regulations will help reduce the testing burden, as districts that used fall-spring testing to report Chapter 1 scores move to a once-a-year cycle. Districts may have other reasons for testing twice a year, however, and one cannot assume they will automatically abandon biannual testing. In addition, the Department's policy to discourage the use of the same test for student selection and pre-testing is an incentive to continue testing in both fall and spring.

USEFULNESS OF FEDERAL INFORMATION

National

Legitimate questions can be raised about how much information Congress really needs to determine whether Chapter 1 is working as intended. Are several million test scores really necessary to produce a chart that shows a single upward trend? Some key

Congressional staff have indicated that they do not receive or make much use of ED's Chapter 1 state summary reports. Others have noted that the reports are not timely.

Aside from putting together the reports required by Congress, ED uses information for internal purposes: targeting states and local districts for monitoring, preparing budget justifications and Congressional testimony, helping to identify exemplary programs, and doing nationally mandated and other studies of Chapter 1. However, ED officials have raised concerns about the quality of the data the Department receives from states. In general, the types of errors to which the data is subject tend to inflate the scores, according to Departmental officials.

State

States use data to monitor compliance, recognize outstanding programs, target technical assistance, and decide state policy.

There is evidence from past evaluations of Chapter 1 that some states have used the Chapter 1 requirements to justify broader reporting systems that are more useful to the state than to the Federal governments.

Local

There is fairly broad agreement that at the local level, Federal Chapter 1 evaluation data is far less useful than other measures of student progress. Building level administrators seldom use NCE scores to modify instruction and have a decided preference for more immediate and student-centered evaluation methods.

ALTERNATIVE ASSESSMENT

Who Is Doing It

A high degree of interest in alternative assessment exists at the local level. Much of the alternative assessment that is occurring with Chapter 1 students is being done at the classroom level and is not considered part of the regular Chapter 1 evaluation system by teachers or district staff. Some districts have begun to institute more formal alternative assessments to measure desired outcomes.

Some districts are doing interesting things with assessment outside of Chapter 1, but because of the Federal framework's emphasis on test scores, they are hesitant to apply it. Some districts prefer to see how alternative assessments work in other arenas before using them in Chapter 1, especially if they are going to be held accountable for progress toward desired outcomes. In other districts, it is the Chapter 1 teachers who are the innovators in alternative assessment.

Given the scope and funding level of Chapter 1, it is remarkable how little effect it has had on encouraging the development of appropriate assessment tools for disadvantaged children.

How Is It Being Used

Alternative assessments are being used in Chapter 1 for diagnosis, student selection, and evaluation of student progress.

State and school districts are using a whole range of alternative assessments and standards to measure desired outcomes; assess progress in early childhood Chapter 1 programs; and inform teacher decisions.

What Forms Are They Using

The entire range: portfolio assessments, developmental checklists, math and science manipulations, holistic writing, end of book tests, reading inventories, teacher-student dialogues, and many others.

What Are the Major Issues

Teachers and other staff need more inservice training to make alternative assessment work well.

Alternative assessments are difficult to score; most of the work must be done by hand. In addition, performance-based assessments are costly.

Alternative assessments for desired outcomes cannot be aggregated beyond the local level.

Some experts have raised questions about whether Chapter 1 students would perform worse on well-executed alternative assessments than on standardized tests. One local person also referred to early evidence in her district that minority children may not do as well on alternative assessment as on criterion-referenced tests.

Some Federal officials feel that alternative assessments are too subjective and cannot produce the hard data needed to compare schools with each other for program improvement. In addition, Federal officials raise the possibility that alternative assessment could be used to make programs look better than they are and ultimately penalize disadvantaged children.

III. CASE STUDIES OF THE IMPACT OF CHAPTER 1 EVALUATION AND ASSESSMENT REQUIREMENTS ON TWO LOCAL SCHOOL DISTRICTS

PURPOSE AND METHODOLOGY OF THE CASE STUDIES

To determine the impact of Chapter 1 evaluation and assessment requirements, the contractor conducted two cases of the School District of Philadelphia and the Portland Public Schools. The case studies sought to answer the following specific questions:

- What effect do the Federal Chapter 1 evaluation and assessment requirements have on local testing procedures and Chapter 1 instructional practices?
- How do local administrators and teachers use the data that is generated to meet Chapter 1 evaluation requirements?
- How do local perceptions of Federal requirements affect the use of alternative assessment strategies?
- To what extent are local districts using alternative assessment in Chapter 1 programs or for Chapter 1 students?
- What suggestions do local practitioners have for improving the Federal Chapter 1 law and regulations concerning evaluation?

To answer these questions, the contractor spent two days in Philadelphia and one day in Portland. Philadelphia and Portland were selected because they both have several thousand Chapter 1 children and large-scale, sophisticated testing programs, but differ in size, geography, demography, problems, and needs. In each site, the contractor met with district Chapter 1 staff, district evaluation staff, and other district leaders such as assistant superintendents, instructional supervisors, and Federal program administrators. In addition, the contractor visited Chapter 1 schools in the two districts and met with

principals and teachers. The contractor also reviewed local evaluation reports and other written documents from the two school systems.

CASE STUDY #1: THE SCHOOL DISTRICT OF PHILADELPHIA

Chapter 1 in Philadelphia

The public schools of Philadelphia enroll over 218,000 students. The majority are African American (63%); the rest are white (23%), Hispanic (9%), and Asian (4%). The School District of Philadelphia faces the same challenges as other big-city districts. The district is divided into seven subdivisions, each overseen by a district superintendent. Among and within these subdivisions exist obvious socioeconomic disparities: to the west of the Board of Education one finds narrow streets of quaint townhouses, farther to the north, blocks of abandoned buildings. Almost half of the students--about 44%--come from families on public assistance. Currently this city of over 1.5 million people faces serious fiscal problems that also threaten the school district's funding base.

As with any large urban district, Chapter 1 is an important component of the overall educational effort. In school year 1990-91, Chapter 1 contributed \$73.7 million to the system's annual operating budget of \$1.17 billion, but the program's influence extends far beyond this 6% share. Over 60% of the city's schools are Chapter 1 schools: 115 elementary schools, 28 middle schools and 17 high schools, a total of 161. Another 50 nonpublic (mostly parochial) schools receive Chapter 1 services. Approximately 2,000 of the district's 20,000 staff are employed by Chapter 1; they provide supplementary instruction in reading and math and support services to about 65,000 participating children in public schools and 7,500 children in nonpublic schools.

A decade ago, the schools of Philadelphia were marked by wide variations among buildings in curriculum and standards, frequent teacher strikes, social promotions from grade to grade, and a lack of public confidence.⁴⁶ Following the arrival of Superintendent Constance Clayton in 1985, the school district took some decisive steps to improve the quality of education. A standardized curriculum, grading guidelines, and promotion policies were instituted; the chain of command was reorganized to give instructional leadership higher priority; and objectives for student achievement, attendance, behavior, and parental involvement were adopted by all schools.

Chapter 1 has undergone reform, too. After passage of the 1988 Hawkins-Stafford legislation, the district created a task force to redesign Chapter 1 in Philadelphia. The result was a total restructuring of the program that concentrated resources in the neediest schools, emphasized prevention over remediation, aligned Chapter 1 more closely with the regular curriculum, expanded the range of services, encouraged school-site management, and more closely monitored student progress. In the new law's schoolwide project provisions--which allow schools with 75% or more poor children to use Chapter 1 funds to upgrade instruction for all children--the district saw an opportunity to refashion Chapter 1 with maximum flexibility. Thus began a districtwide thrust to encourage all qualifying buildings to take advantage of the schoolwide option.

Philadelphia's Testing Program

Since 1984, the school district has administered the Philadelphia City-Wide Test (CWT) to all students in grades K-12. (Before that, the city had used the

Comprehensive Test of Basic Skills (CTBS), and prior to that, the California Achievement Test.) The CWT test was developed by CTB-McGraw Hill and is tailored specifically to the district's curriculum. The CWT consists of both norm-referenced items from CTBS Form U, and criterion-referenced items from the CTB item bank and the Philadelphia school district staff; the two types of items fulfill different functions. (It is not uncommon for test publishers to promote and develop this type of "two-in-one" test keyed to specific local curriculum needs.)

Student scores on the norm-referenced items are to be used for Chapter 1 student selection and evaluation and for comparing Philadelphia's students with the rest of the country. The criterion-reference items are used to determine whether students have mastered the objectives of the Philadelphia curriculum. Each school receives a detailed computer printout showing the percentage of criterion-reference items each student answered correctly, grouped by skills, by grade, and by classroom. Principals and teachers can use these profiles to evaluate strengths and weaknesses at the student, classroom, and school levels, and to adjust instruction accordingly. Parents also receive a printout of their child's profile along with the report card.

Different versions of the CWT are administered twice a year, in January and June. Both testing cycles play a role in Chapter 1. The January test scores are used to determine which students are eligible. The June test scores are compared with the previous June's scores to calculate national percentile ranks and NCE gains for Federal evaluation and program improvement purposes. (The Department of Education

discourages LEAs from using the same test score for student selection and formal Chapter 1 evaluation purposes.⁴⁷⁾

To qualify for Chapter 1 in Philadelphia, a public school child must score at or below the 40th percentile in grades 1 through 4; and at or below the 25th percentile in grades 5 through 12. Thus, a child who scored at the 23rd percentile in January 1991 could be selected to participate in Chapter 1 for 1991-92. That child's June 1991 score would become the pre-test score, before services are provided, and the June 1992 score, the post-test.

The CWT serves other purposes besides fulfilling Chapter 1 requirements and giving individual schools achievement information. Schools' average scores are published in the newspaper, and although the district does not base funding decisions on them, these rankings shape public opinion and serve as one basis for accountability. Individual student scores are also used to determine entry into some of the city's competitive magnet schools.

The school district is facing the fact that it must soon renorm the CWT. Renorming in any big city becomes somewhat of a delicate political situation, because when a test is renormed, the students scores look lower. While this makes no practical difference for Chapter 1 evaluation, because the new scores can be equated to be comparable to past scores, it suggests the extent to which tests have symbolic value to the public and the district's leadership.

The city administers other standardized tests. In March kindergartners take the Stanford Early School Achievement Test (SESAT), a commercial paper-and-pencil test

in which the teacher reads the questions and the children select answers from pictures. Scores from this test are used to determine eligibility for first grade Chapter 1 programs. In March all students in grades 3, 5, and 8 take the state-mandated Test of Essential Learning and Literacy Skills; allocation of special state funds are based on scores from this test (which is about to be discontinued). Several times during the year, students in grades K through 8 also take a math skills tests, called the Philadelphia Mathematical Evaluation Test (P-MET), to determine whether they have mastered the objectives of the math curriculum up to that point. School principals may give other tests for particular purposes, and teacher-made tests are, of course, a part of the classroom routine.

According to school staff, the direct annual cost to the district of all testing is about \$2 million, of which about \$1.3 million goes to McGraw Hill to administer the biannual CWT testing cycles and the remainder pays for district testing personnel and other local costs. About \$1 million more comes out of the LEA's Chapter 1 budget for personnel and other evaluation costs directly attributable to Chapter 1. None of these figures include indirect costs to the schools, such as lost instructional time (four days per year, in the case of the CWT) and diverted staff time.⁴⁸

As the above description of evaluation in Philadelphia makes clear, Federal Chapter 1 requirements have influenced district decisions about testing instruments, testing frequency, and testing schedules. According to Steve Davidoff, the Chapter 1 evaluation director, Chapter 1 is "the backbone of the LEA evaluation effort." James Lytle, former district evaluation director and now a district superintendent, noted that prior to 1965, the city did little testing and that much of the growth in standardized

testing is attributable to Chapter 1. Were it not for Chapter 1, staff said, there would surely still be standardized testing in the district--the public has come to expect it--but perhaps not as often and not at every grade.

A Pioneer in Schoolwide Projects

The authority for schoolwide projects--under which schools with high poverty may serve all children, regardless of eligibility--had been part of Federal law since 1978, but was seldom used because districts had to match Federal dollars in these projects with state and local funds. Philadelphia was one of the few districts in the country operating schoolwide projects before 1988, and along with other urban districts had lobbied Congress to remove the matching requirements. Congress obliged in 1988, at the same time inserting a new requirement: after three years of operation, schoolwide projects must be able to demonstrate that their Chapter 1-eligible children are making higher NCE gains on average than either their counterparts in the district's other Chapter 1 schools or than Chapter 1 children in the same school before adoption of the schoolwide approach. If the school cannot show increased achievement, then they have to go back to serving only eligible children with Chapter 1 funds. As an alternative to higher NCEs, secondary level schoolwide projects can demonstrate lower dropout rates, increased retention rates, or higher graduation rates.

This year Philadelphia is operating 91 schoolwide projects, and next year will have 103--an enormous increase over the 11 projects operated before 1988, in fact more than half the Chapter 1 schools in the city and more than any other system in the country. A visit to the Philadelphia School District makes clear that the district is staking its

reputation on proving that the schoolwide concept can work. The schoolwides are administered by an office called "Priority One/Schoolwide Projects." The district has adopted several policy changes aimed at making schoolwide projects its primary instructional mode. To enable more schools to meet the 75% poverty threshold, the district changed the poverty criteria used to qualify schools for Chapter 1, from AFDC counts alone to a combination of AFDC counts, Refugee Assistance Act eligibility, free and reduced school lunch counts (at the elementary and middle school level) and family food stamp eligibility (at the high school level). By next year, every school that is eligible will be declared a schoolwide project. Using the rationale that these, the poorest schools in the city, have the greatest needs, the district is putting roughly \$1,000 extra dollars per pupil into the schoolwides, over and above what other Chapter 1 schools receive. In addition, the district has shifted some principals to strengthen the leadership of the schoolwides.

The schoolwide approach has wrought many changes in participating schools. The pullout model that characterized Philadelphia's Chapter 1 program before 1988 and is still popular in many of the non-schoolwide projects has been largely abandoned, although some schoolwides still group students with special needs for remedial or intensified instruction. Schoolwide projects are generally using their Chapter 1 funds to reduce class size, hire specialists in reading and math, and in some cases add school-community coordinators who serves as links with the home. The additional per pupil funding helps pay for a program support teacher in every schoolwide project; this person provides staff development and helps classroom teachers implement the project's goals

and monitor student progress. Chapter 1 also funds support staff, so that teachers and other instructional personnel can meet regularly in teams to discuss school goals.

Interviews revealed a general feeling among principals and staff of the schoolwides that the concept has rejuvenated their schools. Every project has set goals for staff development, achievement in reading and math, school climate, and parent involvement, and many have reorganized their staffing and curriculum from the ground up. For example, Dr. Marty Eilberg, principal of Taylor Elementary, spoke enthusiastically of the changes the new approach has brought about in his school, including the use of whole language instruction, cooperative learning, and the teaching of math with manipulatives. Although he is eligible to retire, Dr. Eilberg said he is staying on to see the project through this, its third year.

In one sense, Philadelphia has more riding on the outcome of Chapter 1 evaluation than any other district in the country. The central office staff has invested a great deal of time, energy, and resources into rallying support for the schoolwide concept throughout the district and providing technical assistance, building by building. Similarly, the principals and teachers in every schoolwide building have put enormous effort into implementation. If at the end of their third years, some of these schools do not meet Federal accountability requirements (which given the number, seems almost inevitable), the principals and teachers do not relish the idea of disbanding the programs they have worked so hard to establish and going back to the "old way"--a separately identifiable Chapter 1 program, in many cases a pullout program, which many feel did not help children as much. Some school staff also fear that the district will take away the extra

resources and that their failure to meet accountability standards will become a brand that will follow them through their careers.

Some of the schoolwides are ending their third year, and the staff in these schools are acutely aware of what is at stake in the June round of testing. As one principal said, "the clock is ticking."⁴⁹

Program Improvement in Philadelphia

The basic standard for determining aggregate performance in the Pennsylvania state program improvement plan is any NCE gain above zero, the minimum allowed under Federal regulations. As required by the Federal framework, however, schools must show substantial progress in both aggregate performance and other desired outcomes to stay out of the program improvement process. The Pennsylvania Department of Education has been more active than perhaps any other state in prescribing the types of desired outcomes LEAs can use for Chapter 1 program improvement purposes. The SEA requires school districts to select three desired outcomes, one for each of the three statutory purposes of Chapter 1--success in the regular program, attainment of grade level proficiency, and achievement in basic and more advanced skills--from a state-developed list of options. For grades 2 through 12, the choices for measuring success in the regular program include report card grades, teacher observations and surveys, writing samples, promotion rates, attendance, criterion-referenced tests, informal reading inventories, or another option specified by the district. The choices for determining grade level proficiency include unit tests, level tests, book levels, course levels, report card data, criterion referenced tests, informal reading

inventories, or another locally-developed option. For the third desired outcome, achievement in basic and more advanced skills, there is only one permissible measure: NCE gains that are at or above the state mean.⁵⁰ The state program improvement plan lists minimum NCE gains for reading and math for each grade, ranging from 9.9 NCEs in grade 2 math to 0.1 NCEs in grade 12 reading. The choices of desired outcomes for preschool, kindergarten, and first grade are the same for each of the three purposes: developmental checklists, criterion-referenced tests, teacher and parent observations, writing samples, informal reading inventories and running records, and documented think-alouds. No NCE gains are required for these programs.

A school is determined to have made substantial progress for program improvement purposes if it meets two out of its three chosen outcomes. (This requirement was phased in; schools that were selected in 1989-90 to begin program improvement in 1990-91 had to meet only one of three outcomes.) Under the Pennsylvania plan, it is possible for a school with average NCE gains that are well above zero but below the state means to be put into program improvement, if it did not meet one of its other two desired outcomes.

There is some incentive, then, for school districts to pick measures for judging the other two outcomes that are more within their direct control than test scores. It is also theoretically possible, although not as likely, for a Pennsylvania school district to base program improvement decisions solely on test scores, using NCE gains to determine aggregate performance and achievement in basic and more advanced skills, and criterion-

reference test scores to determine success in the regular program and grade level proficiency.

To measure success in the regular program, Philadelphia has chosen criterion-referenced scores on the CWT for grades K through 8, and increased attendance for grades 9 through 12. To measure grade level attainment, the city uses report card data in grades 2 through 8 and dropout rates in grades 9 through 12. The decision to use grades was based partly on the fact that the city already had in place a computerized, standardized grading system and also on the correlation between grades and promotion, retention, and graduation rates.⁵¹ Not all administrators and teachers seemed satisfied with the reliance on grades. A greater concern expressed by school personnel, however, was that the Federal law and the state interpretation of it have made desired outcomes "another hurdle," rather than an encouragement to explore alternative ways of assessing progress. Several said that the state rules on desired outcomes were too restrictive.

Program improvement began formally in Philadelphia in June 1989, when 46 schools were cited for program improvement based on their June CWT scores. School year 1989-90 was the program improvement planning year for these schools. (The regulations allow a year for planning, with full implementation the following year.) By January of 1990, these schools had begun to put their plans in place. In June of 1990, 25 of this initial group of schools tested out and did not have to implement their plans; another 53 schools were identified for program improvement, with school year 1990-91 the planning year and 1991-92 the implementation year. It is likely that this June's scores will relieve some of these schools of the obligation to fully implement their

plans, while another new group comes into the process. No school has yet reached the joint planning, or "state receivership" stage, as some local officials view it. This pattern of schools exiting and entering during various stages of the process seems to be typical of how program improvement works at the local level.

The changes proposed by the schools in their program improvement plans vary widely, but focus mainly on the grades and subjects that showed negative NCE gains--sometimes to the detriment of the overall Chapter 1 program, according to some principals. Some administrators are convinced that because of anomalies in testing, negative NCE gains⁵² do not always indicate where the greatest needs of the Chapter 1 children lie.

School staff tend to see program improvement as a "scarlet letter," and some are concerned that this identification could negatively affect their careers, according to the district's committee of instructional interventionists. (The instructional interventionists provide instructional leadership to all the schools their assigned sector, so they are in daily contact with school staff.)

Chapter 1 Student Selection in Philadelphia

Philadelphia's procedures for Chapter 1 student eligibility and selection illustrate both the significance of test scores and the influence of Federal audits on local interpretation of Federal policy. The law and regulations require districts to uniformly apply "objective criteria . . . which include written or oral testing instruments" in selecting children. The Federal framework does not require norm-referenced tests for student selection, and in fact permits the use of other criteria, such as teacher assessments,

performance-based assessments, and criterion-referenced tests. However, it has been longstanding practice in Chapter 1--a practice so routine it is scarcely questioned--for districts to use a cutoff percentile score on a norm-referenced test as the primary determinant, and then to make exceptions using other educational criteria. The degree to which other criteria are used varies from district to district. In Philadelphia, schools select for services those children with the greatest educational needs--in other words, those with the lowest test scores--from a computerized list of all children scoring below the cutoff. In exceptional cases, schools can select children above the cutoff who show serious needs according to other measures, but to do so they must follow strict districtwide guidelines and submit an "exceptional case form." This form is retained by the central office as documentation for Federal program monitors and auditors, who, according to Dr. Bob Magliano, the director of categorical programs and government relations, nearly always scrutinize student selection practices. Over the years, Philadelphia, like most large cities, has been involved in major audit disputes with the U.S. Department of Education, and this cautiousness about student selection is just one example of how Federal audits have affected local Chapter 1 practices.

District personnel are aware that this heavy reliance on a single percentile score has its drawbacks. Both Mr. Davidoff and Mr. Lytle noted that children whose scores hover near the cutoff have a tendency to slip in and out of the program.

Usefulness of Chapter 1 Evaluation Data

Central office personnel were more inclined than school level staff to acknowledge the benefits and efficiency of using norm-referenced tests to collect broad-

brush information about fundamental skills and to give Federal policymakers a national picture of Chapter 1 student achievement. But district professionals questioned the need for the Federal government to collect millions of test scores to determine how Chapter 1 is working. Katherine Conner, program manager of Priority One/Schoolwide Projects, and Mr. Davidoff and others felt that a well-constructed sample could do the job as well or better.

Most of the staff, from the central office on down, expressed reservations about the usefulness of norm-referenced tests at the individual student level. Most principals and teachers interviewed felt that the norm-referenced scores on the City-Wide Test were only minimally helpful at the school building level, although some principals saw value in knowing where their schools and students stood in relationship to others. As one principal noted, the percentile scores help tell her where her students stand "in the marketplace, because that is where they ultimately compete."⁵³ Another principal found some utility in the national percentile ranks because they were "purely objective," but said that he used them in conjunction with many other types of data.⁵⁴

The criterion-referenced profiles that the district provides were viewed as more productive by school principals. Most principals interviewed said that the profiles were helpful diagnostic tools that they used to monitor teachers, pinpoint problem areas, assign specialist teachers and support staff, select issues for staff development, pick out students for tutoring, and see how they stand in comparison with the rest of the district. Some gave examples of how instruction had been modified as a result of trends that showed up in the CWT profiles. At the same time, most principals stressed that they did

not rely on the tests as the sole indicator. A variety of indicators informed their decisions, including many subjective ones, such as teacher judgment.

Teachers on the whole seemed to have less faith in test scores of either the norm-referenced or criterion-reference type, but this does not mean they disregarded them. Some teachers said they used the student profiles on the criterion-referenced items, in conjunction with other measures, to corroborate their own judgments about where students were weakest or discover trends they may have overlooked. Even the teachers who were skeptical about standardized tests said that they paid attention to the scores, if for no other reason than knowing that they, their students, and their schools would be judged by them. In general, teachers said that put more faith in assessments of student performance that they had more control over and that gave them information on a daily basis, such as reading levels, chapter tests, and their own observations.

Impact of Chapter 1 Evaluation on Instruction

Philadelphia school people expressed the usual concerns and complaints about standardized tests: they are inappropriate for young children, traumatizing for some children, not attuned to different learning styles, culturally biased, error prone, and inadequate or misleading about individual student performance. Common concerns were also raised about the validity of NCEs in Chapter 1: they are highly technical and hard to explain; are based on questionable assumptions about how much growth a disadvantaged child would make without Chapter 1; do not capture progress that shows up in other indicators; and can be used inappropriately to label children.

Of greater relevance, though, are the Philadelphia staff's comments about how Chapter 1 requirements influence the way they do their jobs. And here, administrators, principals, and teachers agree that the 1988 program improvement and schoolwide project accountability requirements have enormously elevated the importance of test scores in Chapter 1 and have unduly reinforced norm-referenced tests as the main way of assessing student performance. As Mr. Davidoff said, NCE gains have become "the ticket to the game."

If interviews are any indication, teachers in Chapter 1 schools "feel the squeeze" from administrators to make sure their students do well on the CWT. One teacher, whom district staff credit as an innovator in portfolio assessments, said she had transferred out of a schoolwide project because the principal made her feel as though she were "in a fishbowl." This same teacher described how dejected teachers feel when their students do poorly on the tests: "If the kids do well on NCEs, it's nothing I've done, but if the kids don't do well, it's everything I've done."

Local district staff almost universally deny "teaching to the test," but will readily admit to using a number of instructional techniques aimed at improving test scores. Local school people tend to draw fine distinctions between teaching to the test and teaching to the curriculum on which the test is based. Allocating classroom time to skills that are assessed in the CWT is acceptable, in their view, because the test is based on the broad spectrum of skills that make up the city's curriculum. As one former teacher said, "I looked at past tests, and thought, 'this is what they have to master,' then made sure they got it." The principal of a schoolwide project explained how her teachers start

early giving teacher-made tests that are similar in format to the CWT. Another principal said that his teachers all know who the "CHIPS" children (Chapter 1 Intensive Pupils) are and understand that these children's scores are the one upon which the schoolwide project will ultimately be judged. If necessary, the principal said, his teachers will add more services, extra homework, and additional help from paraprofessional staff for these children.

Chapter 1 evaluation also appears to have a limiting effect on instruction. Because tests are so compartmentalized and focused on lower level skills, "teachers can teach just ten things and get good test results," according to an instructional interventionist. Instructional interventionists elaborated further, saying that testing sometimes makes teachers afraid to try new ideas that might improve instruction, like curriculum integration, cooperative learning, and active learning. This group conceded, however, that this is in part the district's fault for promoting a standardized curriculum and stressing the importance of the CWT as a companion.

Principals and teachers also expressed some frustration with the amount of time spent on testing and tracking test data. Even so, say the instructional interventionists, the level of complaints has lessened over the years, not because the amount of time spent on testing has decreased, but because "by now testing is a given."

Concerns about Program Improvement

If one strong theme emerged from the case study, it is that Philadelphia district administrators and principals seem convinced that norm-referenced test scores do not always identify for program improvement those schools with the lowest achievement or

the weakest programs. According to Mr. Davidoff, roughly 30% of the schools in program improvement are there due to measurement phenomena. For example, a school whose average NCE reading score rose from 15 to 16 would not be targeted for program improvement, while a school that declined from 44 to 43 would. To test their suspicions about the capriciousness of the process, the staff analyzed the schools targeted for program improvement in 1989-90. They found that schools with above-average pre-test scores were more likely to post negative NCE gains than schools with low pre-test scores. The study uncovered the following trends:

- Of the 27 elementary schools with negative NCE gains in reading, 20 had average pre-test scores that ranked at or above the district median.
- Of the 18 schools with negative NCE gains in math, 14 had pre-test scores at or above the median.⁵⁵

The analysis also concluded that some schools are in program improvement simply because they had one high-scoring year--a personal best--that was hard to top, although that is what the law requires. As Lionel Lauer, Associate Superintendent for Instruction said, "You can only squeeze so many leaps in NCE gains year after year."

Data from the Hackett School, a K-5 schoolwide project, illustrate how test score phenomena can put a school into program improvement. Although the school has a 96% low-income enrollment based on AFDC data and school lunch counts, achievement scores on the whole have increased significantly between 1988 and 1990. Hackett had the highest average Chapter 1 NCE reading scores in the city in 1989 and the highest math scores in 1989 and 1990. In addition, Hackett ranked first, second, or third in a number of other indicators during the years 1988 through 1990, including average daily

attendance; least number of suspensions; reading, math, and science grades; number of students meeting promotion criteria; percent of students reading on grade level; and percent of all items correct on the CWT. In June 1990, Hackett was identified for program improvement because of declines in CWT math scores in second grade (-1.37 NCEs), third grade (-4.42), and fifth grade (-12.22). The fifth grade average NCEs in particular, which were based on the 1989 and 1990 scores from the nine students for which the district had pre-test and post-test data, dropped from 49.11 NCEs to 36.89. (Schools do not always have pre-test and post-test data for all Chapter 1 children, primarily because of student mobility.)

As a result of these scores, Hackett's program improvement plan focuses on strengthening basic computational skills for all students, as well as providing enrichment and remediation for grades 3 through 5. According to the plan, "all classroom teachers will devote time to drills on basic facts"--an indication that the school intends to target its curriculum more on the type of lower-level math skills that are the primary focus of standardized, norm-referenced math tests. Jim Yoa, Hackett's principal, remains unconvinced that math is the greatest area of need, since average math scores for Chapter 1 eligible students are still higher than average reading scores. (The concern that NCE scores were driving resources toward areas that were not necessarily the biggest problem was echoed by principals of other schools in program improvement.) Mr. Yoa is optimistic, however, that the June CWT math scores will be high enough to excuse the school from full implementation.⁵⁶

Principals and district staff pointed out several other factors that could put a school into program improvement.

- Several school people cited high mobility rates--50%, 60%, 75% turnover in individual schools--as a factor that could negatively affect NCE scores and put them into program improvement. As long as students stay within the district, their scores follow them. A transient student may have a pre-test and some Chapter 1 services in one school, then transfer in mid-year to another school, where his post-test score contributes to the average. (It should be noted that the ED Chapter 1 policy manual does not require students who have enrolled in the program "well into the year" to be included in the evaluation.)⁵⁷ There was a widespread feeling voiced by school staff that they wanted to be held accountable only for instruction provided on their site. Mobility can affect program improvement in another way. If large numbers of students in a school move outside the district, then the school could be judged based on average NCEs from the small number of children for whom the district has both pre-scores and post-scores; in this situation, one child's score can make a significant contribution to the average.
- Even in situations without high mobility, a few very low scores can affect the average. One school professional said, for example, that as a result of state policies, children who would have been put in special education in the past are now being placed in Chapter 1, and that some of these children score very low on standardized tests.
- With the city's depressed economy, the demography of some schools has changed rapidly, so that each incoming Chapter 1 group is more disadvantaged than the last. For schools in deteriorating neighborhoods, just staying at the same performance level was an accomplishment, many staff stressed.
- One mediocre teacher, inexperienced teacher, or frequently absent teacher can affect the scores for a whole grade and skew the results for the entire school, according to some Philadelphia principals. One principal felt her school was in program improvement mostly because one class had a series of substitutes for the better part of a year.
- The principal of a schoolwide project that was also in program improvement cited the language problems of many children in his 80% Hispanic school as a contributing factor. Although not considered limited English proficient by the district's definition, many of these children still

• speak Spanish in the home and have trouble with the reading sections of norm-referenced tests, the principal said.

- Because of limited city, state and Federal resources, children are not getting all the social services they need, and this is affecting their achievement.

Many of these same conditions were cited as reasons why a schoolwide project could fall short of Federal accountability standard.

Philadelphia school personnel also expressed frustration with the Federal requirements and state policies for individual student program improvement. Federal law and regulations require districts to make a thorough assessment of need and revise instruction for students who have been in the program two or more years and have not shown improvement. The Chapter 1 policy manual defines improvement as positive reading and math NCEs. District staff felt that this reliance on test scores was inappropriate and overlooked other important indicators of student progress.

Concerns about Schoolwide Project Accountability

District staff expressed a number of concerns about the Federal requirements for evaluating schoolwide projects.

First, there was some confusion in Philadelphia about certain U.S. Department of Education policies. District officials were unclear as to whether ED would allow them to average achievement in the schoolwides across the three years, or whether they must consider only the achievement made during the last of the three years.⁵⁸ There had also been some uncertainty about whether to include preschool, kindergarten, and first grade students in the evaluation of schoolwide projects, but recently the district found out that this was required.

Another concern pertains to comparison groups. Some district staff questioned the fairness of comparing schoolwide projects with non-schoolwide projects--one of two ways of determining progress under the Federal law--because the schoolwide projects are by definition the poorest schools in the district and because, in Philadelphia, the non-schoolwide group is smaller. Moreover, according to Ms. Conner, because the Chapter 1 population in the non-schoolwide projects is a smaller percentage of the total enrollment, there would more regression toward the mean than in schoolwide projects, where the Chapter 1 eligible group tends to comprise about two-thirds of the enrollment. Federal law also allows schoolwide projects to compare the current achievement of Chapter 1 eligible children with the achievement of Chapter 1 children during the three prior years. School officials said the demographic changes that have occurred in some schools could make this unfair.

Many school staff felt that it would be fairer to judge schoolwide projects according to the performance of all the children in the school, not just the Chapter 1 eligible group.

A final concern was that some of the curriculum changes that schools are making may take more than three years to show up in test scores. Ms. Conner cited an example of a school that was commended as an exemplary program by the U.S. Department of Education--the first schoolwide project so recognized--in which it took four years for the NCE scores to improve, even though success had begun to show up earlier on other indicators.

Alternative Assessment

Philadelphia teachers are using many different forms of alternative assessment for Chapter 1 children, but because of the preeminence of test scores in the Federal framework, they tend not to view these assessments as part of the formal Chapter 1 evaluation process. The case study found teachers using portfolio assessment, running reading records, detailed observation reports, writing assessments, checklists, math and science performances and a range of others.

Although there is no districtwide policy on alternative assessment, the district has incorporated alternative assessment into some of its programs. The district's desired outcomes for pre-kindergarten programs requires schools to use developmental checklists to assess progress of Chapter 1 children at this level. As another example, Ms. Cecilia Cannon, an instructional support teacher in the Office of School-Age Programs, has put together a package of assessments for the Benchmark program, a Chapter 1-funded effort in non-schoolwide buildings that uses a thematic teaching approach and small class sizes. The assessments include the Miscue Analysis, which evaluates students' reading skills through close teacher observation of read-alouds; reader response questions to be used in conjunction with children's literature; teacher observation forms; written and oral comprehension questionnaires; problem solving procedures for small groups; and a student self-evaluation of reading skills and attitudes. Of course, the children in the Benchmark program must go through regular Chapter 1 testing, as well.

As part of Superintendent Clayton's districtwide reform efforts, all schools were required to develop school improvement objectives. Some schools are evaluating

progress toward these objectives using alternative assessments, such as learning logs, teacher-student "dialogues," informal reading inventories, and math and science performances. Principals interviewed said they found these measures to be more in synch with what their teachers were teaching than standardized tests. As one principal said, "This is the data we have more control over and can do something about [by modifying instruction] right away."

District staff feel strongly that the Chapter 1 requirements have a chilling effect on alternative assessment. Even though the law does not prohibit alternative assessment, and the policy manual gives examples of desired outcomes based other indicators, there is a pervasive feeling that NCEs are all that really matter. As one district Chapter 1 evaluator noted, "The teachers say, 'tell us about alternative assessment when you can convince us that NCEs are not all that people care about.'" The committee of instructional interventionists noted that even parents tend to value the "one number" in the student profile more than the descriptive and qualitative information teachers provide with report cards. Some district staff spoke wistfully of the positive influence the Federal Chapter 1 law could have if it gave credence or encouragement to alternative assessments.

This is not to say that the school personnel wanted to use alternative assessment as a replacement for standardized tests. As one principal pointed out, there is a need for some objective data that is not teacher-determined; otherwise, there are no "checks" on teacher who are poor judges or have low expectations. In general, staff felt a

combination of alternatives assessments, standardized tests, and other indicators was the best approach to evaluation.

Suggestions for Change

Some of the Philadelphia district professionals have given considerable thought to ways of improving the Federal evaluation and assessment requirements. As would be expected, some of their recommendations are aimed at making the program easier to implement in their district. Overall, the recommendations would have applicability to many large cities.

First, Mr. Davidoff, Ms. Conner and others recommended separating national evaluation purposes from local. They felt that a well-constructed sample could provide sufficient national data, thereby eliminating the need for districts to test every student every year. This sampling could be overseen by the Federal government and administered through the states and local agencies. A large district like Philadelphia, they felt, would be able to construct a reasonable random sample of students, classrooms and schools. Implicit in this recommendation is an assumption that some other means besides norm-referenced tests would be allowed for the annual program improvement and student improvement reviews.

There seemed to be some support among teachers for initiating testing later, perhaps in grade three. While the Federal law does not require pre-kindergarten through first grade to be tested for Chapter 1, a district like Philadelphia that uses a spring-spring annual cycle in effect must test first graders to get a pre-score for second grade services, or do a special round of tests in the fall of second grade for this purpose.

A number of people also wanted more flexibility in the measures that could be used to judge program improvement. The general suggestion was to judge substantial progress using either desired outcomes or achievement scores, instead of both. They also wanted the ability to develop their own desired outcomes; as long as the standard was the same throughout the district, they did not see a problem. Another recommendation was to change the law so that program improvement could be calculated using average achievement over a period of years, instead of year by year.

For schoolwide projects, several people suggested that accountability be based on the average performance of all the children who have been in the school for a service year, not just Chapter 1 eligible children. Other recommendations for schoolwide changes included:

- Allowing a grace period or planning year for schools that did not meet accountability requirements, similar to the planning year in program improvement; during this time they could make plans to modify the instructional program and also try to bring up their test scores.
- Adding a state case by case appeal process for program improvement; if the state determined the decline in test scores was due to uncontrollable circumstances, such as budget cuts, it could declare the results invalid.
- Giving the state authority to determine that a revision of a schoolwide model had more promise for helping Chapter 1 eligible children than a return to the discrete Chapter 1 program, in schools that did not meet accountability standards.
- Eliminating from program improvement and accountability decisions those children who had been in a school for less than a program year.

Observations

Because of Philadelphia's investment in schoolwide projects, Chapter 1 test scores have taken on a unique significance. While Philadelphia, like other districts, is

concerned about program improvement requirements, it is the schoolwide project accountability measures that will make the greatest impact when the June test cycle is completed. The first group of schools will be judged then, and there are likely to be long hours of number-crunching as the staff tries out the allowable comparisons to qualify as many as possible for continuation.

It should not be forgotten that the intent of the accountability measures was to ensure that in commingling Chapter 1 funds with other funds, schoolwide projects did not dilute the intensity or effectiveness of services to Chapter 1 eligible children, the most disadvantaged children in the school. If some schools do not meet the accountability requirements, it could be a legitimate indication that the special needs of Chapter 1 children have become secondary to the overall goals of reform in some schools. For this reason, one must look carefully at some of the legislative recommendations for schoolwide projects, to be sure they do not loosen accountability to the point that the Chapter 1 eligible children are treated just like any other. For example, if continuation of schoolwide projects were based on the achievement of all children in the school, instead just the Chapter 1 eligible children, it would be possible for the rest of the school to score high enough that a decline or stagnation in the scores of Chapter 1 children would be masked.

There is nevertheless an irony in the fact that while the purpose of the schoolwide project authority is to allow greater flexibility to upgrade instruction in poor schools, the method of determining a project's success is so inflexible: an NCE gain. Questions must be raised about whether some other indicators should be brought into the process.

Although the staff in these projects stress the innovative instruction that is occurring in their schools, the significance of test scores is omnipresent. If, as a result of June testing, a high percentage of schoolwide projects do not make the necessary gains, one cannot help but wonder what effect this will have on the other schoolwides. Will they revise their programs to emphasize more the skills that are tested by the norm-referenced items, even if it goes against the spirit of their reformed instructional program? (It should be pointed out, however, that principals and teachers do not know which items on the CWT are the norm-referenced ones.) Will the schools that fall out go back to a pullout program or use an in-class model that complies with Chapter 1 anti-supplanting provisions?

With respect to both schoolwide projects and program improvement, the Philadelphia case study provides some telling illustrations about the effects of relying on a single, imperfect measure to make important decisions. When the stakes are so high, as they are in program improvement and schoolwide projects, there is a natural temptation to try to attribute low performance to outside factors, and some of the reasons cited--substitute teachers, a large NCE loss in a single grade, low scores of students at the bottom of the curve--may say more about school policies than about the form of measurement. However, district staff gave several examples of how the capriciousness of measurement can fail to pinpoint the schools with the greatest problems or the greatest areas of need within schools. These examples speak to the need for Congress to reexamine test scores as the primary determinant of school

progress and are a persuasive argument for allowing more flexibility, or at least an opportunity to appeal when legitimate uncontrollable circumstances affect the outcome.

Similarly, one must ask how realistic it is for a school to keep raising its test scores year after year. As any teacher will testify, the makeup of a class can vary greatly from year to year, and a drop in NCE gains need not signify that something is wrong with instruction. The way that program improvement is set up, before too many years it is possible that the majority of Chapter 1 schools in the country will have gone through the process, and a substantial percentage more than once. While it is good for schools to reexamine their programs periodically, some other indicators, in addition to test scores, might help them more accurately judge their real progress. Again, one must be careful not to loosen the strings too much, so that the needs of children are not forgotten. If unambitious desired outcomes were used to excuse schools from program improvement, that would defeat the purpose, to ensure that Chapter 1 students receive quality instruction.

The view of alternative assessment as being something separate from Chapter 1 is a key finding. As the Philadelphia experience illustrates, the Federal framework, intentionally or not, serves to inhibit alternative assessment; in this sense, Chapter 1 requirements are an anchor against the tide of reform in assessment.

A final observation: the Philadelphia testing program shows just how intertwined Federal and local evaluations have become, to the point that it is difficult to sort out what factors are behind which testing decisions. While the district staff attribute much of the growth of testing to Chapter 1, it is clear that as regards the CWT, the district has its

own testing goals and uses apart from Chapter 1, and that Chapter 1 testing and local accountability testing expanded concurrently. With a test that is so linked to the curriculum and such a regular part of school life, one cannot be sure to what extent Philadelphia or any large district would cut back on testing in the absence of Chapter 1 requirements. To some extent, Chapter 1 could serve as a cover under which any local school district could carry out its own testing purposes while pointing the finger at the Feds. Even so, the Philadelphia case study illustrates how using a single system for national and local evaluation needs may not do either justice and raises a key issue about whether the needs should be met through different systems.

CASE STUDY #2: PORTLAND PUBLIC SCHOOLS

Chapter 1 in Portland

Portland, Oregon, is a city of over 400,000 situated near the confluence of the Willamette and Columbia Rivers. One does not have to travel far beyond the city limits to find breathtaking scenery, evergreen forests, and fishing streams; even so, the city confronts many of the same social and economic challenges as other urban areas. The Portland Public Schools enroll 54,904 pupils in 90 buildings. The majority are European American (72.1%); the rest are African-American (15.2%), Asian (7.6%), Hispanic (3.0%), and American Indian (2.1%).

The district takes pride in the overall rising trend in test scores that has occurred since 1980-81, under Superintendent Matthew Prophet, and particular pride in the extent to which African-American and other minority students have made progress in closing

the achievement gap. For a city its size, Portland has a relatively low dropout rate--about 8%--with an average dropout rate among African American students that is slightly below the districtwide average.

Sixty-one of the city's public schools are Chapter 1 schools, including 44 elementary schools, 11 middle schools, and 6 high schools. Schools are chosen using a poverty criteria that was recently changed from a combination of AFDC and free and reduced school lunch counts to school lunch counts alone.

Chapter 1 in Portland supports supplementary reading and math instruction for about 8,000 students, in either pullout or in-class settings, depending on the school. Three of the elementary schools have been designated as schoolwide projects, and one of these schoolwides is finishing up its third year, the year of reckoning.

Students are eligible for Chapter 1 in the 1991-1992 school year if they score below the 33rd percentile on the city's Portland Achievement Levels Test (PALT); previously the cutoff score had been the 25th percentile. Although the change expands the pool of eligibles, the number of newly eligible students served will depend on how much new Chapter 1 money the district receives. The Chapter 1 budget for schools year 1990-91 is \$5,868,353.

Portland's Testing Program

Portland has a highly sophisticated research and evaluation (R &E) department with a national reputation for being on the cutting edge of assessment research and technology. The R & E department employs 29 full-time equivalent staff positions, including 14 evaluation specialists, two of whom devote much time to Chapter 1. The centerpiece of Portland's testing program is the PALT.

Portland began using standardized tests regularly in the early 1960s. After finding that no commercial instrument was aligned closely enough with its curriculum, the district set about developing its own test to fulfill the functions of both a norm-referenced and a criterion-referenced test. The result was the PALT.

It is clear the district staff do not think of the PALT as just another standardized test. Since 1977, it has been the primary local evaluation instrument, but over the years it has been greatly expanded and refined. The PALT uses all locally-developed items, aligned to the district's curriculum in reading and math. A PALT in language usage (punctuation and other writing conventions) was retired at the end of 1989-90.

The PALT are functional tests, a certain kind of what testing experts call narrow-range tests, which means that a student takes only those items from the overall test that are slightly above his or her prior achievement level, as determined by past performance. To maximize the accuracy of the measurement, test forms and levels are assigned so that each student is expected to get approximately half of the items correct.

Using a technique called item response scaling, all the items in the item bank and test scores are calibrated by difficulty using the Rasch Unit (RIT) scale. The range for items is 120 to 280; the range for scores is more typically 150 to 250. From the item bank, the R & E staff produces up to eight distinct forms of per grade, each with a different level of difficulty. The resulting student scores can be related to a common scale, such as an NCE scale; or broken down by curriculum goal, grade, and subject, to indicate whether students have mastered curriculum objectives. Computer-generated reports of individual student test results are given to schools, classroom teachers, and

parents. Aggregate reports are given to cluster and program administrators, as well as district policymakers and the community.

PALT reading and math tests are administered twice a year, in October and April, to every student in grades 3 through 8. If, by the spring of eighth grade, a student has met or surpassed the district's minimum basic skills standards in reading and math (required to graduate from high school with regular diplomas), then formal testing ceases. Students who have not met or surpassed these standards by the time they enter high school are tested periodically with the paper and pencil Graduation Standards Test (given each spring districtwide) or on the Computer Adaptive Reporting and Testing (CARAT) system, as explained below.

Two PALTs in (a) science concepts and process and (b) science topics were given in spring 1991 for the second time in 3rd grade, the third time in 5th grade, and the fourth time in 8th grade. Some students in 4th, 6th, and 7th grades took developmental versions that are used to calibrate new items.

Since 1977, PALT has been the instrument for evaluating Chapter 1 in grades 3 through 8. (High school Chapter 1 students take the Stanford Achievement Test; second graders take the SRA listening test.)

Portland is especially proud of its vast bank of several thousand test items. The PALTs in reading and math are updated continuously, throughout every six-year curriculum adoption cycle, and new items are added to the bank after each developmental testing period. Approximately 80 new reading and math forms are produced for each season.

To make its tests available across the country, Portland founded a consortium called the Northwest Evaluation Association, which also includes other districts in Oregon and Washington. Districts that are interested in adapting Portland's test to their own uses can buy the whole system, including the item bank and computerized software, for \$15,000. The software enables a district to feed in its curricular goals and get back a test geared to its needs. About 15 districts in the country use PALT-based tests, including some other large cities. Funds from test marketing go into the district's operating budget, but they are a minor source of the district's revenues.

The Portland staff is also proud of the other strides it has made in testing design and technology. For example, the math tests have been redesigned so that students can use calculators on many sections, and the reading tests incorporate passages from children's literature, consistent with the whole language approach used in many elementary schools. Both reading and math tests have been revised to place more emphasis on higher order skills, according to the staff.

Computers are used to develop tests and manage the item bank, as well for test scoring, data management, and reporting. Portland's computer system also helps staff track Chapter 1 children who move to other schools within the district.

Computers also play a role in the administration of tests. With the CARAT system, school staff can administer PALTs on-site at any time. Microcomputers programmed with CARAT software are now in all Chapter 1 schools and by next fall will be in every school.

Although the regular fall-spring administrations of the PALTs are via paper-and-pencil, the CARAT tests serve a variety of other purposes: assessing entering students, grouping students, and testing high school students who have not met graduation requirements. CARAT also allows teachers and administrators to quickly obtain existing PALT data on any designated group of students in the school.

According to Dr. Jane Arends, assistant director for testing and evaluation, there are schools where a CARAT test is given every day. Diana Thompson, a specialist in the CARAT system, said that Chapter 1 teachers and students are among the most active users of CARAT; often the machine sits in the Chapter 1 classroom. If a school has several CARAT stations, each with a computer, monitor and Bernouli drive, then several students can take a CARAT test at the same time.

A CARAT test begins with randomly-chosen test items linked to the subject and purpose selected by the teacher. Depending on how the student performs, the computer selects each next item from the bank according the student's response on the prior item. CARAT tests take about 20 minutes. Each test is unique, with items balanced for content and variety.

CARAT has several advantages, said Portland staff. It is faster and more accurate than a paper-and-pencil test; allows schools to gauge student performance with fewer items; has a lower error rate; is better liked by students; and opens no opportunity for students to cheat. The R & E staff said they have researched whether the computer is a factor in student performance on the tests and found that it is not.

Portland students take other standardized tests in addition to the PALT. Five years ago, the district began a local direct writing assessment in eighth grade, and added fourth grade two years ago. The program was suspended for 1990-91 because the state implemented an almost identical assessment in grades 3, 5, 8, and 11. Students in those same four grades throughout the state also took tests in reading, literature, math, listening, and study skills. Portland was given permission to substitute PALT data for 3rd, 5th, and 8th graders for the reading, literature, and math tests, eliminating the need to "double test" students. Permission was granted because the state tests were developed from the same item bank Portland uses. (Chapter 1 students in other districts in the state took the state assessments in reading and math in addition to the commercial tests their districts used for Chapter 1 pre- and post-testing.)

Portland must still give state tests in listening skills, study skills, and writing to these three grades, and all the tests to 11th graders. The Portland district staff raised questions and concerns about the new, mandated statewide assessments. The district's major objection is that schools and districts will be compared on the basis of student achievement levels, rather than gains. Portland's "Principles of Test Result Reporting," like Chapter 1 regulations, make student achievement gain the only proper basis for comparisons and accountability. In particular, they feared that this state assessment marked the beginning of "high stakes testing" in Oregon, and that the state might eventually use results to allocate state dollars. If this happens, said Dr. Walter Hathaway, director of R & E, it could compromise the integrity of the testing process, as "people find a way to get the numbers they need."

After sixteen years of front-end developmental costs, the annual cost of testing, including staff, operations, and equipment, is now less than one-half of 1% of the district's budget, according to Dr. Hathaway.

Impact of Chapter 1 on Testing

For many years, Portland has used the PALT results for Chapter 1 reporting, with NCEs generated from the "national norms" of the California Achievement Test, to which PALT has been equated.

The national evaluation standards that the U.S. Department of Education published in response to the 1988 Hawkins-Stafford law created two dilemmas for the Portland testing program. First, the standards required all LEAs to use nationally-normed tests so that data could be aggregated; states and local districts that wanted to use their own equatable tests had to apply to ED for special permission. After conversation's with the state and ED, Portland requested ED permission to use local norms for generating NCEs, the only local district to make such a request thus far.

Portland wants to use PALT norms for several reasons:

- PALT information, say the district's R & E experts, is superior in terms of measurement characteristics and curriculum alignment to that obtained from any commercially available test.
- PALT reports are more comprehensive and sophisticated than any commercially available.
- Local norms will always be current and thus more valuable to those who seek to understand the strengths and weaknesses of current Chapter 1 programs.
- PPS uses local norms for all other norm-referenced reports of test results, so consistency between reports for Chapter 1 and other reports of aggregate test results would be achieved.

Portland's request is still under review by the Department; in the meantime, Portland is continuing to use the PALT for Chapter 1. If the Department decides that local norms do, in fact, produce valid and reliable NCE scores and approves the request, it could be a boon to Portland's reputation and budget by creating new demand for PALT-based tests in other districts.

A second problem Portland has with the national Chapter 1 evaluation standards relates to the requirement that LEAs use an annual evaluation cycle. This means that Portland can no longer report Chapter 1 gains using fall-spring test scores. PPS prefers the fall-spring cycle because the districts uses it for all other reports of student achievement gains. Moreover, say district staff, with an annual cycle, the high turnover rates in Chapter 1 school reduce the number of clear and intact groups of students (those who are tested at both points and in the same school) to unacceptably low levels.

After trying out a spring-spring cycle last year, Portland decided to go with a fall-fall cycle. Portland staff particularly did not like spring-spring cycle because it meant testing first graders in the spring in order to have a pre-test for second grade Chapter 1 students. In general Portland staff believe standardized tests are inappropriate for young children. (In fact, there is a general distaste among district and building-level staff for testing before third grade.) In addition, ED discourages using the same scores for student selection and pre-testing. Since the Portland staff preferred not to use fall scores to select students because enrollments are not yet settled, they are using spring scores for student selection and fall for pre- and post-testing.

As to whether Chapter 1 increased the amount of testing in the district, the staff said that the district probably would not give as many tests were it not for Chapter 1. With two full-time equivalent positions devoted exclusively to Chapter 1, it is evident that Chapter 1 testing affects staff time. Other R & E staff also spend unreimbursed time on Chapter 1 evaluation, said Dr. Arends; for example, during March and April of 1991, as spring testing neared, the R & E staff spent about one-fourth of its total staff time on Chapter 1.

Program Improvement in Portland

According to the Oregon Department of Education standard, any school with NCE gains under + 1 are targeted for program improvement. The state department of education has proposed raising the standards to +5 NCEs for elementary schools and +3 for secondary. Along with other districts in the State, Portland opposes this change.

They argue that:

- Virtually all schools would have to participate in program improvement, and there are not sufficient resources to go around.
- Schools do not like inflexible regulations or the stigma of mandated improvement efforts.
- It is difficult to explain "aggregate NCE" to parents and other lay people, without adding the complication of explaining an arbitrary increase in the standard.

In 1989, seven Portland schools were identified for program improvement, on the basis of spring test scores. These schools had the 1989-90 year to plan, during which four of them tested out. The remaining three are implementing program improvement this year. In the meantime, another group of 12 schools had been identified; they are

planning this year. The Oregon SEA has instituted a process it calls voluntary program improvement for school districts that are just above the minimum. An additional nine Portland schools are developing voluntary plans this year.

The confusion over which testing cycle to use made program improvement difficult to implement, as the district switched first to a spring-spring cycle and then to a fall-fall cycle for formal evaluation purposes. Depending on which cycle was used, a whole different set of schools were in or out of program improvement, although a subset showed up on all cycles.

Another problem with program improvement, according to staff interviewed, is high mobility: students who come into a school in November have no pre-test; those who leave before April have no post-test. Some Portland schools have been identified for program improvement on the basis of very small clear and intact groups of Chapter 1 students for whom they have both pre- and post-test scores. According to Dr. Carolyn Moilanen, the Chapter 1 staff person who oversees program improvement, one school failed to meet the criterion of a 1.0 NCE gain in reading because small gains by two 5th-graders did not offset the high gain from one 4th grader. Beach School principal Michael Verbout noted that his school, with its 40% mobility rate, had small clear and intact numbers of Chapter 1 groups with the cycle used and would not have been required to participate in program improvement had another cycle (and a different clear and intact group) been used.

Dr. Moilanen said that although school staff generally felt chagrined about being in program improvement, the attitudes seemed to be growing more positive. Schools

that accept (even like) program improvement are those which have either linked it to schoolwide reform or just need some fine tuning. For example, she said, in one school the clear problem was its fragmented schedule, which was relatively easy to change.

Approximately 30 low-performing schools per year undertake a major school improvement effort supported with district funds. Thus, schools targeted for Chapter 1 program improvement may view it as an unnecessary duplication of effort, district staff noted.

The schools in program improvement have proposed a variety of alterations, according to Dr. Moilanen, with some undertaking radical curriculum changes. The most positive outcome of program improvement, in her opinion, is the way it has brought about better coordination between Chapter 1 and the regular program.

Interviews with building level staff suggest that at least in some cases, the program improvement process has targeted genuine needs for reform in Chapter 1. For example, the principal of one school that went into voluntary program improvement because of average math scores just barely above the minimum said that math has been slighted in her school, and she hopes that program improvement will convey the message that "math must be sacred." Similarly, another school cited for program improvement because of its math NCEs has one Chapter 1 teacher who is certified in reading but teaches both reading and math. This school has an extensive list of changes it is examining for next year, including better coordination between Chapter 1 and regular program staff, teacher team meetings, more integrated curriculum, parenting and child development classes for Chapter 1 parents, an ungraded approach in K through 3, and a set block time for all

pullout programs. The principal of this school remarked that the special state funds available for schools in program improvement create a strange, inverse incentive: "If you do well, you work yourself out of money," he observed.

The SEA has prescribed no desired outcomes except NCE scores, leaving the decision of whether to write desired outcomes up to LEAs. The state and the Federal regulations do require desired outcomes for early childhood programs, however, and Portland is wrestling with which measures to use. According to Dr. Moilanen, there is no single good indicator because the preschool and kindergarten curriculum is not standard throughout the district. In math, the district is considering clinical assessments using manipulatives, which fits with the math programs in most early grades in the district. For reading, the decision is much harder, since programs range from traditional to innovative. Part of the problem is with the language of the Chapter 1 law, according to Dr. Art Burke, an evaluation specialist who works directly with Chapter 1 and other externally-funded programs. Dr. Burke noted that the purposes stated in the law, such as 'grade-level proficiency' and 'basic and more advanced skills,' are difficult to interpret for children in the early years. For grades 2 through 12, the district staff were undecided whether they will set desired outcomes other than aggregate performance.

The Impact of Chapter 1 Evaluation on Instruction

Because Portland uses the same test for Chapter 1 and district purposes, it is difficult to sort out how much testing is attributable to Chapter 1 alone. Principals and teachers interviewed complained about the amount of time spent on testing in general and about the disruption testing makes in the academic schedule. According to one

principal, the state testing and PALT testing together take up nine half-days; the SRA test for second graders takes two or three more. This principal said she did not understand why fall testing was necessary and expressed concerns that the President's proposals for education will lead to more standardized testing.

The district R & E staff countered that no matter how much testing is reduced, there will always be complaints. They said that ironically, building level personnel have generally said that if Chapter 1 testing is necessary, they would rather have all students tested than have Chapter 1 students pulled out for testing, which the staff said the schools found disruptive.

The PALT tests appear to have affected instruction, insofar as teachers and principals admit to using a range of strategies to prepare students for the tests. Most of these strategies were related to teaching skills in the district curriculum they know will be on the test. Teachers and district staff both acknowledged that principals sometimes pressure teachers to make sure their students perform better on the tests.

As regards the impact of computerized testing, the staff of the one school said that while they liked CARAT and thought it was helpful, they do not have the time and staff to use it to its fullest extent.

Usefulness of Chapter 1 Evaluation Information

The Portland district staff maintain that the primary emphasis of PALT testing is to support instructional decision-making. To see whether this occurs, the district has done studies of the use of PALT data. The studies have found that teachers and principals in Portland are far more likely than their counterparts elsewhere to make use

of test results. One study found that 97% of Portland principals, 88% of directors of instruction, and 85% of teachers used PALT evaluation in some way. The largest group, 40%, were routine users who were comfortable and knowledgeable with the data but did not use it much to enhance instruction; 21% were "enhanced users." (National studies, according to Dr. Hathaway, generally show less than 20% of teachers using publisher's standardized test results in any way.)

Portland case studies also found that teachers and principals used the information differently. Principals, who tended to use the data more frequently, said it helped them with such tasks as setting goals, placing students, reviewing student growth, regrouping students, evaluating curriculum, and communicating with parents and teachers. Teachers were most likely to use data for communicating with parents, reviewing individual student growth, and placing students. Even so, the studies found, building level staff preferred other forms of assessment to standardized tests for making decisions about instruction. PALT tests ranked below performance assessment, oral questioning, teacher-made tests, class assignments, textbook tests, and other teachers' opinions as data used by teachers in instructional planning.

In response to findings from its research on teacher use of PALT results, the districts research and evaluation department has taken a leadership role nationally in developing and applying various forms of performance-based assessment.

Interviews with building-level staff found principals and teachers agreeing that while the NCE scores were seldom used to modify instruction, the PALT printouts of student mastery of curricular goals were used for many purposes. District policy

prohibits the use of PALT scores for teacher evaluations; however, as one principal pointed out, score reports are analyzed in planning for staff development and other forms of instructional improvement. Another principal said that in his school, the curriculum director meets individually with teachers to go over PALT results. Further, he said, while the PALT was a "superb survey test" that told him how many children needed which type of skills, it, like any other standardized test, did not diagnose individual student needs with any precision.

In response to requests from many teachers and principals for tests with greater diagnostic precision, a diagnostic level of the CARAT system is under development.

In summary, the building level staff had mixed feelings about the PALT, which had as much to do with standardized testing in general as with the particular instrument. As one principal said, "PALT is good as tests go, but it still leaves some things to be desired." Building level staff said they relied the most on teacher-made tests, which they felt more reliably assessed student performance.

Concerns about Chapter 1 Evaluation

District staff, principals, and teachers in Portland expressed a number of concerns about the Chapter 1 evaluation requirements.

The district staff raised several technical concerns about the use of the NCE as the standard. As Dr. Arends noted, the NCE is a "purely mathematical concept; its link to substantive issues is several steps removed from reality." Further, she explained, the difference between +1 NCE and +2 NCE means very little in terms of a child's performance. Dr. Burke said that "the assumption that the aggregated NCE scores index

program effectiveness is a 'fantasy,' and basing conclusions upon only two points in time is replete with methodological problems." He suggested that the Federal government "get off the commensurability-drawn train and encourage local accountability."

Another problem with the NCE, said the staff, is that it is difficult to explain to the school board, principals, teachers and the public. The district staff said they would rather be allowed to use a standard tied to their own curriculum, such as the Rasch Units that they use to calibrate the PALT.

These feelings about NCEs lead central office staff to view Chapter 1 evaluation requirements as somewhat of a compliance exercise. Because of the district's computerized data system, the staff feel they can provide whatever statistics the Department wants without too much burden, then go about their own local evaluation concerns. As Dr. Arends said, "The goofiness of some of the regulations has not suppressed inventiveness, scholarship, development, and research to create an outstanding system for measuring student performance." In addition, she said, it has not held back teachers from trying some "very interesting classroom assessment."

Principals and teachers were unswaying in their beliefs that the NCE scores do not reflect the true performance of Chapter 1 children. The following specific comments about testing Chapter 1 children were voiced:

- "Chapter 1 children are assessed to death, which means they are constantly experiencing failure."
- "Elementary children get off track filling in the bubbles."
- "I have seen second graders weep."

- "Lots of our Chapter 1 children come from bad home situations. We see growth that is unmeasured by tests. Just getting these children to want to come to school is an achievement."

In general, the building level staff wanted more flexibility to use diverse forms of assessment in Chapter 1.

Concerns about Chapter 1 Student Selection

Portland has traditionally used teacher judgment in combination with cutoff scores to determine which children are selected for Chapter 1. As a result of a recent visit by ED monitors, Portland is rethinking its policy. The monitors complained that teacher judgment was not being applied consistently across buildings. Although the district Chapter 1 staff provides forms for teachers to use, Norman Bengel, the Chapter 1 director, says that the staff is looking at some alternative measures, such as checklists, that would make the process more systematic. In general, Mr. Bengel said, Chapter 1 in Portland has become more standardized across buildings as a result of Federal and state requirements and audits.

Building level staff also raised some specific complaints about the use of test scores to select students for Chapter 1. One principal pointed to the erratic decisions that sometimes occur when a difference of one or two percentiles means that a child is in or out of the program. Because teachers must go through paperwork to make exceptions, they sometimes do not bother with it, said one teacher.

Alternative Assessment

To respond to the press for more authentic assessment, Portland has initiated several districtwide efforts. As noted above, the district has suspended its direct writing

assessment in grades 4 and 8 because of the virtually identical statewide program in other grades. Under both the local and the state assessment, students write and revised papers over the course of three days. Two raters mark each paper on several specific "traits."

In reading, the district's curriculum division has developed a collection of classroom-based assessments for grades 1 through 8. This booklet, which staff say has won the favor of teachers, includes a primary spelling checklist, a reading strategies checklist, a reading habits survey, reading journal prompts, a retelling checklist, and a comprehension inventory.

In addition, the district is implementing portfolio assessments in language arts and math in a cluster of eight elementary schools. Teachers are collecting student writing samples, using the spelling checklist, keeping records of students' independent reading, and sampling text comprehension. Standard criteria and procedures for portfolio assessment in language arts grades 1 through 12 are under development.

At the classroom level, Chapter 1 teachers reported using a number of alternative assessment approaches with Chapter 1 children on a more informal basis. One Chapter 1 teacher has developed a first grade checklist that covers letter recognition, number recognition, alphabet letters, counting, geometric shape recognition, name writing, and sound-say skills.

The district R & E staff is visionary about the potential of technology to make performance-based assessment more cost-effective and objective. Dr. Hathaway ultimately envisions the day when alternative assessment can be carried out with

computer and video technology and instruction and assessment will become a seamless whole. CARAT, for instance, "could someday evolve into a normal part of classroom instruction, and the district could 'skim' student performance information to serve districtwide accountability purposes. Students, for example, could conduct computer-simulated science experiments in class, which would also yield data for science program assessment."

Staff cited several obstacles to the wider use of alternative assessment. Chapter 1 is one impediment, according to Dr. Burke, inasmuch as the time spent on standardized testing is a disincentive to use more alternative assessment. In addition, the stakes tied to program improvement create a wariness about using alternative assessments to measure other desired outcomes, when one is unsure what they will show.

Another obstacle is cost. Dr. Hathaway felt that "a great deal more money would have to be spent on research and development to make alternative assessment practical for large scale as well as classroom assessment purposes. Paying trained people to make consistent and comparable about children's performance costs \$20 to \$30 per hour." In general, Dr. Hathaway estimated, the cost of alternative assessment is 100 times greater than regular testing, and if one cut corners, the quality of the assessment is compromised.

In addition, Portland district personnel said, a massive staff development effort is necessary to train teachers and other staff in proper alternative assessment techniques.

Finally, the technology has to be improved to make information from alternative assessments "portable and storable," said Dr. Arends. For example, she said, the

Reading Recovery program, a nationally-known program, requires extensive and precise diagnostic workups of individual first graders. Dr. Arends would like to see technology developed to reduce the recording burden for teacher and simultaneously yield anonymous and aggregatable student data for program evaluations.

Suggestions for Change

Portland staff had the following recommendations for changes in the Federal framework:

- Begin mandatory standardized testing later, perhaps in third- or fourth grade.
- Allow schools to use alternative assessment for program improvement decisions.
- Develop a new model for local accountability that is independent of "commensurability" and aggregation, and permit local districts to use less formal, more frequent measures that looked at student growth across several points in time.
- Make Federal funds available to support local research and development of alternative assessments, locally-developed tests that serve formative evaluation needs, and computer applications.
- Permit LEAs to use their own evaluation systems as long as they comply with the standards of the Joint Committee on Standards for Educational Evaluation. (These standards address such issues as utility, feasibility, propriety, and accuracy of evaluations.)
- Allow LEAs to use whatever reasonable testing cycle they wish. Allow formal Chapter 1 evaluation every two years as a local option.
- Eliminate the rule that the same data cannot be used for both student selection and pre-test.

Observations

Over the years, Portland's research and evaluation effort has evolved into a highly sophisticated operation that is in the vanguard of developments in assessment. One might argue that this sophistication in and of itself gives impetus to more testing, and that the district's investment of time and resources in refining the PALT makes the district less likely to consider other alternatives. The interest of the district staff in pursuing alternative assessments suggests that this is not the case, although it seems clear the district assumes that PALT will continue to be a major part of the picture. If ED accepts the PALT as an alternative to commercially published tests and other districts express interest, it is hard to predict exactly what this will mean for Portland.

Responsibility for marketing PALT, CARAT, and other "products" of the district's measurement research and development efforts is delegated to the Northwest Evaluation Association. It is possible that the PALT may provide districts who tailor it to their own needs with better formative information and more curriculum alignment than commercial publishers' tests. Even so, it would not diminish the amount of standardized testing in Chapter 1.

It should be noted, too, that the PPS district staff view the PALT somewhat differently from the building level staff. The Research and Evaluation Department believes that the PALTs are as good as (and in many ways superior to) other standardized testing systems for both formative and summative evaluation purposes. From the limited amount of information one can glean from a brief visit, it seems the

building staff are more likely to feel that although the PALT is more useful than most standardized tests, it is still another test, and in that sense an imperfect measurement.

Like many districts in the nation and the State, Portland has been somewhat slow in implementing certain provisions of the 1988 amendments to Chapter 1, such as the 12 month testing cycle and desired outcomes. This type of delay is not uncommon, as first the U.S. Department of Education, and then Oregon and other State educational agencies take time debating, developing, and communicating new procedures and interpretations to all districts. According to PPS staff, the state Board of Practitioners are still debating and addressing several new requirements, and a document describing the new guidelines and procedures had yet to be issued as of this writing.

CONCLUSIONS FROM CASE STUDIES

- In large urban districts, broad-scale, sophisticated testing programs have evolved concurrently with the expansion of Chapter 1. For this reason, it is hard, if not impossible, to determine how much testing is occurring because of the presence of Chapter 1. To some extent, the Federal requirements provide a cover for testing that meets local needs more than Federal.
- If Chapter 1 national evaluation requirements were to disappear tomorrow, it might affect the frequency of testing in urban areas but it is unlikely it would reduce testing by a substantial amount.
- District level staff to some extent use test data as a checkup of school performance. Principals to some extent use it as a gauge of teacher performance.
- Testing in urban areas serves purposes that are as much symbolic as practical. The public in local districts has become accustomed to relying on numerical indicators as indicators of district progress. Superintendents have come to see their value in symbolizing that the district is achievement-oriented. It is the existence of the numbers themselves and not their validity that fulfills these functions.

- Chapter 1 evaluation data based on NCE scores has little real formative value at the school, classroom, and student levels.
- The limitations of nationally-normed tests make them a very imperfect and perhaps inappropriate measure for assessing the progress of individual children and schools.
- As the new program improvement provisions are being implemented, it is becoming more clear that the use of test scores are not necessarily targeting the schools with the greatest problems or the greatest problems within schools. However, there is a tendency to blame external factors for why a school is in program improvement, even when there are legitimate instructional reasons.
- The implementation of the 1988 amendments' schoolwide project and program improvement requirements have raised the stakes placed on evaluation and on test scores enormously and may be influencing instruction in ways that are counterproductive to the goals of Chapter 1. Although one of the intents of the 1988 law was to move away from drill-type programs for disadvantaged children and toward the teaching of higher order skills, the reliance on test scores to judge progress may be having just the opposite effect.
- Although the desired outcomes provisions of the law have the potential to open up new opportunities for alternative assessment, these opportunities cannot be fulfilled the way the Federal framework is structured and the way some states are interpreting it.
- While many teachers use alternative assessments with Chapter 1 students as part of daily instruction, they do not view this as a formal part of Chapter 1 evaluation.
- Chapter 1 requirements may be an obstacle to expanded use of alternative assessment.
- It is natural for school district administrators to want to loosen the strings in Federal programs. While some additional flexibility about evaluation may be entirely warranted, the Federal government must be sure that in providing more flexibility that meeting the needs of Chapter 1 children, and not local convenience, is the foremost goal.

PART IV. FEDERAL POLICY OPTIONS

INTRODUCTION

The time is ripe for a reexamination of the Federal requirements for evaluation and assessment that have governed Chapter 1 practice since implementation of TIERS in the late 1970s. The provisions of the 1988 amendments that turned Chapter 1 evaluation from a routine compliance exercise to a high-stakes statistical enterprise have served to highlight some of the weaknesses of relying on norm-referenced test scores to fulfill both national summary and local instructional needs. There is a high interest in alternative assessment methods among states and local school districts, along with some promising models.

In rethinking Chapter 1 evaluation, Congress should make a comprehensive review of the evaluation, student selection, program improvement, and schoolwide accountability requirements as related components of Chapter 1 assessment, not as discrete sections. When reauthorizing Chapter 1, Congress should consider a fresh rewrite of the aforementioned provisions, perhaps grouping all assessment and accountability sections together, without being wed to language simply because it has been in the law a long time. Congress should look closely at sections 1019 and 1021 and make an effort to eliminate any contradictory provisions and vague language that are working at odds with overall intent of law. As part of this review, Congress be attentive to and consider rewriting language that unwittingly endorses norm-referenced testing even when that is not the intent.

KEY QUESTIONS

As Congress rethinks the evaluation and assessment provisions, Congress should carefully consider the following questions:

- What does Congress need to know about Chapter 1? Is Congress getting what it needs? What are the Department of Education's data needs, and are they different from Congress's needs?
- Is nationally aggregated data doing what it was intended to do? Do aggregates of aggregates of averages of gain/loss scores present a meaningful national picture of how well Chapter 1 is working? Are the assumptions on which the NCE is based valid enough to instill trust in the cumulative data? Are high technical standards followed by all the entities--from the school on up--that contribute to the aggregate?
- Are the types of national data that Congress receives in the ED state summary reports useful? Timely? How does Congress use them? Are they justifiable in terms of the cost, time and effort of obtaining several million test scores? To what extent is the value symbolic--to give the appearance of accountability--rather than substantive? Can symbolic values be fulfilled with one type of information, and the substantive with another?
- Is it realistic to serve national, state and local needs with the same information system based on norm-referenced standardized tests? Is it necessary to use the same evaluation system for the national evaluation reporting and for the annual program improvement?
- How do the data needs of state and local agencies differ from those of the Federal government and each other?
- What sort of data meet Federal policymakers criteria for objectivity? Are test scores the only type?

POLICY OPTIONS

The following options are the result of many conversations with state and local educators and other Chapter 1 personnel. Some have been derived from case studies and original observations, some from the recommendations of people who have given

thought to these issues. The options may be considered in combination or separately, as some of them are not mutually exclusive.

Major Changes

Separating National Needs from State and Local

National Data -- Congress could consider separating national evaluation purposes from state and local and developing different systems for each. Congress could obtain a national picture of Chapter 1 through well-constructed, periodic (not necessarily annual) sampling methods that adhere to the highest technical standards, thereby eliminating the need for districts to test every student every year. This sampling could be overseen by the Federal government and administered through the states. Congress could establish specifications for this national sample, much as it has done for the National Assessment of Educational Progress or the various mandated studies of Chapter 1, to collect the precise types of information it needs to make policy decisions.

State and local data -- If the national purposes of the evaluation were separated from the local purposes, it would create a whole new way of looking at Chapter 1 assessment. It would free SEAs and LEAs to use a variety of assessment instruments that better reflected state and local goals. If desired, Congress could allow each state to set its own standards for evaluation, using TIERS models at its discretion, that would govern LEAs in that state. Under this option, TIERS would become a vehicle for setting high voluntary standards and not a system for collating and aggregating state data. If desired, Congress could also require ED to collect and transmit to Congress state reports (which might look very different from each other and would not be aggregatable, but

might nevertheless provide information useful for legislators, Federal program managers, and researchers interested in particular states).

Aggregation

Congress could consider discontinuing the Federal requirement for LEA evaluations to produce aggregatable data. TIERS could be eliminated, made optional at an SEA's or LEA's discretion, or reduced to collecting basic demographic data and/or state summary reports of state or locally designed evaluations and program improvement. The requirement for national evaluation standards could be dropped, and ED could be directed to eliminate all regulatory requirements based on the concept of aggregation.

Local Evaluation

The Federal law could set some very basic standards for LEA evaluations (objective, consistent for all schools in district, conforming with professionally-recognized technical standards, etc.) and allow LEAs to design their own evaluations that met local needs, with the help of technical assistance centers.

Expanded TACs

The technical assistance centers (or another appropriate entity of Congress' choice) could become regional field centers with responsibilities for evaluation. Much as GAO or ED's monitoring division sends out people to audit programs, these centers could be charged with evaluating a specific number of projects annually, using the highest standards and professional staff.

Separating Annual Program Review from Regular Evaluation

If Congress desired, it could require LEAs to conduct periodic evaluations (presumably less than annually) in accordance with whatever standards Congress saw fit, and then permit LEAs to develop more flexible procedures for the annual review that determines which schools are in program improvement. If desired, states could be given authority to establish statewide program improvement standards, much as they do now, except with more flexibility to use instruments other than norm-referenced tests. LEAs could still establish desired outcomes, but would have more flexibility to use alternative assessments in measuring progress toward these outcomes and fulfilling local formative evaluation purposes.

Substantive Changes

Less Frequent Local Evaluations

As an alternative to the changes discussed above, Congress could clarify that local evaluations that meet national standards need only be conducted once every two or three years. The SEA timelines could be made consistent with the local timelines. The statute could clarify that annual review for program improvement purposes could be based on assessments other than the national evaluation and not necessarily on norm-referenced tests. Rather than charging states with designing appropriate samples, Congress could require all states to be on the same timeline, so that an aggregated national picture would be provided just once every two or three years. While the data might not be as timely as it is now, it could be adequate for reauthorization cycles and other national purposes.

More Flexible National Standards

Congress could allow states to choose three or four measures, one of which is a written test: for example, norm-referenced tests, state-developed criterion-referenced tests, secondary school dropout rates. Data would not be aggregated but would be reported by state.

Expanding Early Childhood Exemption

Congress could exempt second and perhaps even third grades from the norm-reference testing requirement.

Program Improvement

Congress could permit substantial progress to be judged using either desired outcomes or achievement scores, instead of both. Under this option, Congress could define with more rigor the types of desired outcomes that would be acceptable and require that the standard be the same throughout the district.

Encouragement of Alternative Assessment

Congress could add language to the law that explicitly encourages the use of alternative assessments for school and student program improvement, measurement of desired outcomes, and local decisions about instruction. For example, Congress could make clear that progress toward desired outcomes and assessments of individual student progress ("student program improvement") can include measures other than NRTs, such as performance assessment and measures of success in regular classroom.

Without an explicit endorsement, it is unlikely SEAs and LEAs will push forward to a very great extent.

Optional Testing Cycles

Consider could allow LEAs to use a fall-spring cycle if they desired. If aggregation is retained, the Department could report fall-spring data separately as it did prior to the 1988 amendments.

Student Eligibility and Pre-Testing

Congress could explicitly allow LEAs to use the same data for determining eligibility and pre-testing. If aggregation is retained, the Department could be directed to widely disseminate procedures for correcting for regression to the mean in these cases, or could direct the TACs to make the necessary adjustments.

Less Substantive Changes

Demographic Data

Congress could delete the requirement for LEAs to submit data on age, since this is widely thought to be not very useful. Depending on how frequently they are used, Congress could also drop requirements to collect information on gender, race, and handicapping conditions of Chapter 1 participants.

Local Conditions

Congress could clarify whether or under what circumstances "local conditions" may be used to exempt schools from program improvement. If exemption is permitted, Congress could clarify that it applies for one year only.

Appeals

Congress could allow LEAs to appeal to the states for waivers for schools that did not meet schoolwide accountability criteria or school program improvement criteria

because of uncontrollable circumstances, which could be spelled out in the law. Schools could be given a one or two year grace period, and then they would have to comply fully.

Schoolwide Projects

Congress could allow a one-year grace period or planning year for schools that did not meet accountability requirements, similar to the planning year in program improvement; during this time they could make plans to modify the schoolwide instructional program and try to improve achievement.

Program Improvement Calculations

For program improvement purposes, Congress could allow aggregate performance to be calculated across three years, instead of year by year, at the local option. In any event, Congress might consider clarifying whether preschool through first grade programs are intended to be included in the calculations of program improvement.

Mobility Allowance

Congress might consider eliminating from program improvement and accountability decisions the achievement scores of children who had been in a school for less than a specified period of time.

Research and Demonstration on Alternative Assessment

Congress could reserve a portion of national Chapter 1 evaluation funds specifically for research, development, and demonstration of alternative assessments that are appropriate for Chapter 1 children, cost effective, and meet quality standards. LEAs, SEAs, institutions of higher education, technical assistance centers and other public and private nonprofit agencies could apply for competitive grants.

Staff Development

Congress could permit funds to be used at the local level (perhaps from the 5% innovation pot) for staff development to help teachers and other appropriate staff learn ways to better assess Chapter 1 children.

Department of Education Option

Regardless of the action Congress takes, ED could make important clarifications to the policy manual so that it is clearer whether each interpretation is an option or a statutory or regulatory requirement.

ACKNOWLEDGEMENTS

The contractor extends special appreciation and thanks to the school districts of Philadelphia and Portland for their cooperation with and contributions to this study. The contractor is particularly appreciative of the time the local administrators, teachers, and other professionals spent in making the site visits productive.

The contractor would also like to thank Mary Jean LeTendre and the staff of the Department of Education's compensatory education programs, as well as the Chapter 1 staff of the Office of Planning, Budget and Evaluation for sharing their time and expertise.

ENDNOTES

1. Of this \$6.2 billion, approximately \$5.5 billion is distributed by formula to local school districts. The remainder is used for three state-administered programs for migrant, handicapped, and neglected and delinquent children and for other specialized programs and activities, such as state administration and Federal technical assistance.
2. Between 1965 and 1981, the program was known as Title I. Its name changed to Chapter 1 in 1981 with enactment of the Education Consolidation and Improvement Act.
3. 20 U.S.C. 2701-2891 (1965), as amended.
4. 34 C.F.R. 200.1-200.89 (May 19, 1989).
5. U.S. Department of Education, Chapter 1 Policy Manual (Washington, DC: April 1990).
6. NCEs range from 1 to 99, with a mean of 50.
7. Mary Kennedy et al., The Effectiveness of Chapter 1 Services (Washington, DC: U.S. Department of Education, 1986), p. E-2.
8. John Chrysostom Dougherty, Report on Changes under Chapter 1 of the Education Consolidation and Improvement Act (Washington, DC: U.S. Government Printing Office, September 1985).
9. Robert M. Stonehill and Curtis L. Groves, "U.S. Department of Education Policies and ESEA Title I Evaluation Utility: Changes in Attitudes, Changes in Platitudes," Educational Evaluation and Policy Analysis, vol. 5, No. 1, Spring 1983, p. 66.
10. Ibid.
11. Prior to enactment of the Department of Education Organization Act in 1979, Title I was administered by the Office of Education within the Department of Health, Education and Welfare.
12. Elizabeth R. Reisner et al., Assessment of the Title I Evaluation and Reporting System (Washington, DC: U.S. Department of Education, 1982), p. 1.
13. Ibid.
14. Ibid., p. 121.
15. Ibid., p. 3.

16. Elizabeth R. Reisner and Ellen Marks, The Federal Administration of Evaluation, Program Improvement, and Technical Assistance under ECIA Chapter 1 (Washington, DC: Policy Studies Associates, 1987), p. 16.
17. Beatrice F. Birman et al., The Current Operations of the Chapter 1 Program (Washington, DC: U.S. Department of Education, 1987), p. 130.
18. Dougherty, op. cit., p. 149.
19. Reisner and Marks, Federal Administration, p. 6.
20. Birman, op. cit., p. 131.
21. Reisner and Marks, Federal Administration, p. ii.
22. The statute was named in honor of retired Congressman Augustus F. Hawkins, former Chairman of the House Committee on Education and Labor; and retired Senator Robert T. Stafford, former ranking Republican and one-time Chairman of the Senate Subcommittee on Education, Arts and Humanities.
23. House Report 100-95, p. 23.
24. Beth Sinclair and Babette Gutmann, A Summary of State Chapter 1 Participation and Achievement Information for 1987-88 (Washington, DC: U.S. Department of Education, 1990), p. 48.
25. House Report 100-95, p. 22.
26. 20 U.S.C. 2722.
27. 20 U.S.C. 2724.
28. Conference Report 100-567, p. 324.
29. P.L. 101-305.
30. Mary Jean LeTendre, personal interview, March 12, 1991.
31. Alan Ginsburg, personal interview, April 2, 1991.
32. Sinclair and Gutmann, op. cit., p. 110.
33. Conference Report 100-567, p. 324.
34. Mary Jean LeTendre, personal interview, March 12, 1991.
35. 20 U.S.C. 2731.

36. 20 U.S.C. 2731.
37. Mary Jean LeTendre, personal interview, March 12, 1991.
38. Senate Report 100-222, p. 16.
39. Alan Ginsburg, personal interview, April 2, 1991.
40. Kennedy, *op. cit.*, pp. 3-4.
41. *Ibid.*, p. 2.
42. *Ibid.*, p. 5.
43. Sinclair and Gutmann, *op. cit.*, p. 160.
44. Howard Essl, telephone interview, May 7, 1991.
45. *Ibid.*
46. Lionel Lauer, Associate Superintendent for Instruction, School District of Philadelphia, unpublished text of speech, Oct. 19, 1988.
47. According to the Chapter 1 policy manual (p. 141), this is because students at the lower end of the scale tend to regress significantly toward the mean. When a group of students is selected from a larger group only on the basis of low pre-test scores, the scores of this lower group, when post-tested on the same instrument, will automatically increase even without any instruction and would therefore show artificially inflated gains. Districts that choose to use the same scores for both purposes must correct for the regression effect.
48. James Lytle, district superintendent, Philadelphia School District, personal interview, April 10, 1991.
49. Brenda Artwell, principal, Belmont School, personal interview, April 10, 1991.
50. James Sheffer, Pennsylvania Chapter 1 state coordinator, telephone interview, March 18, 1991.
51. Katherine Conner, program manager for Priority One/Schoolwide Projects, personal interview, April 10, 1991.
52. In the Chapter 1 program, a decline in NCEs relative to the norming population is referred to as a negative gain rather than a loss, because it does not signify that the child's achievement did not increase; rather that it did not increase at a rate at least as great as the norming group.

53. Sybil Waldman, principal, Comegys School, personal interview, April 11, 1991.
54. Dr. Marty Eilberg, principal, Taylor School, personal interview, April 11, 1991.
55. Katherine Conner, Priority One, School District of Philadelphia, telephone conversation, May 3, 1991.
56. Jim Yoa, principal, Hackett School, personal interview, April 11, 1991.
57. U.S. Department of Education, Chapter 1 Policy Manual (Washington, DC: April 1990).
58. The Department policy is to allow districts to use either method, according to Bill Lobosco, U.S. Department of Education, Compensatory Education Programs division in a telephone conversation on May 2, 1991.